Improving Computational Complexity in Statistical Models with Second-Order Information

Tongzheng Ren[◇,‡] Jiacheng Zhuo[◇] Sujay Sanghavi[†] Nhat Ho[♭,‡]

Department of Computer Science, University of Texas at Austin[◇],
Department of Statistics and Data Sciences, University of Texas at Austin[♭]
Department of Electrical and Computer Engineering, University of Texas at Austin[†],
February 8, 2022

Abstract

It is known that when the statistical models are singular, i.e., the Fisher information matrix at the true parameter is degenerate, the fixed step-size gradient descent algorithm takes polynomial number of steps in terms of the sample size n to converge to a final statistical radius around the true parameter, which can be unsatisfactory for the application. To further improve that computational complexity, we consider the utilization of the second-order information in the design of optimization algorithms. Specifically, we study the normalized gradient descent (NormGD) algorithm for solving parameter estimation in parametric statistical models, which is a variant of gradient descent algorithm whose step size is scaled by the maximum eigenvalue of the Hessian matrix of the empirical loss function of statistical models. When the population loss function, i.e., the limit of the empirical loss function when n goes to infinity, is homogeneous in all directions, we demonstrate that the NormGD iterates reach a final statistical radius around the true parameter after a logarithmic number of iterations in terms of n. Therefore, for fixed dimension d, the NormGD algorithm achieves the optimal overall computational complexity $\mathcal{O}(n)$ to reach the final statistical radius. This computational complexity is cheaper than that of the fixed step-size gradient descent algorithm, which is of the order $\mathcal{O}(n^{\tau})$ for some $\tau > 1$, to reach the same statistical radius. We illustrate our general theory under two statistical models: generalized linear models and mixture models, and experimental results support our prediction with general theory.

1 Introduction

Gradient descent (GD) algorithm has been one of the most well-known and broadly used (first-order) optimization methods for approximating the true parameter for parametric statistical models [21, 3, 19]. In unconstrained parameter settings, it is used to solve for optimal solutions $\widehat{\theta}_n$ of the following *sample loss* function:

$$\min_{\theta \in \mathbb{R}^d} f_n(\theta), \tag{1}$$

where n is the sample size of i.i.d. data $X_1, X_2, \ldots, X_n$ generated from the underlying distribution P_{θ^*}. Here, θ^* is the true but unknown parameter.

When the step size of the gradient descent algorithm is fixed, which we refer to as *fixed-step size gradient descent*, the behaviors of GD iterates for solving the empirical loss function f_n can be analyzed via defining the corresponding *population loss* function

$$\min_{\theta \in \mathbb{R}^d} f(\theta), \tag{2}$$

‡ Correspondence to: Tongzheng Ren (tongzheng@utexas.edu) and Nhat Ho (minhnhat@utexas.edu).

where $f(\theta) := \mathbb{E}[f_n(\theta)]$ and the outer expectation is taken with respect to the i.i.d. data $X_1, X_2, \ldots, X_n$. An important insight here is that the statistical and computational complexities of fixed-step size sample GD iterates $\theta^t_{n,\text{GD}}$ are determined by the singularity of Hessian matrix of the population loss function f at θ^*. In particular, when the Hessian matrix of f at θ^* is non-singular, i.e., $\nabla^2 f(\theta^*) \succ 0$, the previous works [1, 11] demonstrate that $\theta^t_{n,\text{GD}}$ converge to a neighborhood of the true parameter θ^* with the optimal statistical radius $\mathcal{O}((d/n)^{1/2})$ after $\mathcal{O}(\log(n/d))$ number of iterations. The logarithmic number of iterations is a direct consequence of the linear convergence of fixed step size GD algorithm for solving the strongly convex population loss function (2). When the Hessian matrix of f at θ^* is singular, i.e., $\det(\nabla^2 f(\theta^*)) = 0$, which we refer to as *singular statistical models*, $\theta^t_{n,\text{GD}}$ can only converge to a neighborhood of θ^* with the statistical radius larger than $\mathcal{O}((d/n)^{1/2})$ and the iteration complexity becomes polynomial in n. In particular, the work of [11] demonstrates that when the optimization rate of fixed-step size population GD iterates for solving population loss function (2) is at the order of $1/t^{1/\alpha'}$ for some $\alpha' > 0$, and the noise magnitude between $\nabla f_n(\theta)$ and $\nabla f(\theta)$ is at the order of $\mathcal{O}(r^{\gamma'}(d/n)^{1/2})$ for some $\alpha' \geq \gamma'$ as long as $\|\theta - \theta^*\| \leq r$, then the statistical rate of fixed-step size sample GD iterates $\|\theta^t_{n,\text{GD}} - \theta^*\|$ is $\mathcal{O}((d/n)^{\frac{1}{2(\alpha'+1-\gamma')}})$ after $\mathcal{O}((n/d)^{\frac{\alpha'}{2(\alpha'+1-\gamma')}})$ number of iterations. Given that the per iteration cost of fixed-step size GD is $\mathcal{O}(nd)$, the total computational complexity of fixed-step size GD for solving singular statistical models is $\mathcal{O}(n^{1+\frac{\alpha'}{2(\alpha'+1-\gamma')}})$ for fixed dimension d, which is much more expensive than the optimal computational complexity $\mathcal{O}(n)$.

Contribution. In this paper, to improve the computational complexity of the fixed-step size GD algorithm, we consider the utilization of the second-order information in the design of optimization algorithms. In particular, we study the statistical guarantee of *normalized gradient descent (NormGD) algorithm*, which is a variant of gradient descent algorithm whose step size is scaled by the maximum eigenvalue of the Hessian matrix of the sample loss function, for solving parameter estimation in parametric statistical models. We demonstrate that we are able to obtain the optimal computational complexity $\mathcal{O}(n)$ for fixed dimension d under several settings of (singular) statistical models. Our results can be summarized as follows:

1. **General theory:** We study the computational and statistical complexities of NormGD iterates when the population loss function is homogeneous in all directions and the stability of first-order and second-order information holds. In particular, when the population loss function f is homogeneous with all fast directions, i.e., it is locally strongly convex and smooth, and the concentration bounds between the gradients and Hessian matrices of the sample and population loss functions are at the order of $\mathcal{O}((d/n)^{1/2})$, then the NormGD iterates reach the final statistical radius $\mathcal{O}((d/n)^{1/2})$ after $\log(n)$ number of iterations. When the function f is homogeneous, which corresponds to singular statistical models, with the fastest and slowest directions are at the order of $\|\theta - \theta^*\|^{\alpha}$ for some $\alpha > 0$, and the concentration bound between Hessian matrices of the sample and population loss functions is $\mathcal{O}(r^{\gamma}(d/n)^{1/2})$ for some $\gamma \geq 0$ and $\alpha \geq \gamma + 1$, then the NormGD iterates converge to a radius $\mathcal{O}((d/n)^{\frac{1}{2(\alpha-\gamma)}}$ within the true parameter after $\log(n)$ number of iterations. Therefore, for fixed dimension d the total computational complexity of NormGD to reach the final statistical radius is at the order of $\mathcal{O}(n\log(n))$, which is cheaper than that of the fixed step size GD, which is of the order of $\mathcal{O}(n^{1+\frac{\alpha}{2(\alpha-\gamma)}})$. Details of these results are in Theorem 1 and Proposition 1.

2. **Examples:** We illustrate the general theory for the statistical guarantee of NormGD

under two popular statistical models: generalized linear models (GLM) and Gaussian mixture models (GMM). For GLM, we consider the settings when the link function $g(r) = r^p$ for $p \in \mathbb{N}$ and $p \geq 2$. We demonstrate that for the strong signal-to-noise regime, namely, when the norm of the true parameter is sufficiently large, the NormGD iterates reach the statistical radius $\mathcal{O}((d/n)^{1/2})$ around the true parameter after $\log(n)$ number of iterations. On the other hand, for the low signal-to-noise regime of these generalized linear models, specifically, we assume the true parameter to be 0, the statistical radius of NormGD updates is $\mathcal{O}((d/n)^{1/2p})$ and it is achieved after $\log(n)$ number of iterations. Moving to the GMM, we specifically consider the symmetric two-component location setting, which has been considered widely to study the statistical behaviors of Expectation-Maximization (EM) algorithm [1, 7]. We demonstrate that the statistical radius of NormGD iterates under strong and low signal-to-noise regimes are respectively $\mathcal{O}((d/n)^{1/2})$ and $\mathcal{O}((d/n)^{1/4})$. Both of these results are obtained after $\log(n)$ number of iterations.

To the best of our knowledge, our results of NormGD in the paper are the first attempt to leverage second-order information to improve the computational complexity of optimization algorithms for solving parameter estimation in statistical models. Furthermore, we wish to remark that there are potentially more efficient algorithms than NormGD by employing more structures of the Hessian matrix, such as using the trace of the Hessian matrix as the scaling factor of the GD algorithm. We leave a detailed development for such direction in future work.

Related works. Recently, Ren et al. [22] proposed using Polyak step size GD algorithm to obtain the optimal computational complexity $\mathcal{O}(n)$ for reaching the final statistical radius in statistical models. They demonstrated that for locally strongly convex and smooth population loss functions, the Polyak step size GD iterates reach the similar statistical radius $\mathcal{O}((d/n)^{1/2})$ as that of the fixed-step size GD with similar iteration complexity $\log(n)$. For the singular statistical settings, when the population loss function satisfies the generalized smoothness and generalized Łojasiewicz property, which are characterized by some constant $\alpha' > 0$, and the deviation bound between the gradients of sample and population loss functions is $\mathcal{O}(r^{\gamma'}(d/n)^{1/2})$ for some $\alpha' \geq \gamma'$, then the statistical rate of Polyak step size GD iterates is $\mathcal{O}((d/n)^{\frac{1}{2(\alpha'+1-\gamma')}})$ after $\mathcal{O}(\log(n))$ number of iterations. Therefore, for fixed dimension d, the total computational complexity of Polyak step size GD algorithm for reaching the final statistical radius is $\mathcal{O}(n)$. Even though this complexity is comparable to that of NormGD algorithm, the Polyak step size GD algorithm requires the knowledge of the optimal value of the sample loss function, i.e., $\min_{\theta \in \mathbb{R}^d} f_n(\theta)$, which is not always simple to estimate.

Organization. The paper is organized as follows. In Section 2 and Appendix A, we provide a general theory for the statistical guarantee of the NormGD algorithm for solving parameter estimation in parametric statistical models when the population loss function is homogeneous. We illustrate the general theory with generalized linear models and mixture models in Section 3. We conclude the paper with a few discussions in Section 4. Finally, proofs of the general theory are in Appendix B while proofs of the examples are in the remaining appendices in the supplementary material.

Notation. For any $n \in \mathbb{N}$, we denote $[n] = \{1, 2, \ldots, n\}$. For any matrix $A \in \mathbb{R}^{d \times d}$, we denote $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ respectively the maximum and minimum eigenvalues of the matrix A. Throughout the paper, $\| \cdot \|$ denotes the ℓ_2 norm of some vector while $\| \cdot \|_{\text{op}}$ denotes the operator norm of some matrix. For any two sequences $\{a_n\}_{n \geq 1}, \{b_n\}_{n \geq 1}$, the notation

$a_n = \mathcal{O}(b_n)$ is equivalent to $a_n \leq C b_n$ for all $n \geq 1$ where C is some universal constant.

2 General Theory of Normalized Gradient Descent

In this section, we provide statistical and computational complexities of NormGD updates for homogeneous settings when all the directions of the population loss function f have similar behaviors. For the inhomogeneous population loss function, to the best of our knowledge, the theories for these settings are only for specific statistical models [6, 27]. The general theory for these settings is challenging and hence we leave this direction for future work. To simplify the ensuing presentation, we denote the NormGD iterates for solving the samples and population losses functions (1) and (2) as follows:

$$\theta_n^{t+1} := F_n(\theta_n^t) = \theta_n^t - \frac{\eta}{\lambda_{\max}(\nabla^2 f_n(\theta_n^t))} \nabla f_n(\theta_n^t),$$
$$\theta^{t+1} := F(\theta^t) = \theta^t - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta^t))} \nabla f(\theta^t).$$

where F_n and F are the sample and population NormGD operators. Furthermore, we call θ_n^t and θ^t as the sample and population NormGD iterates respectively.

For the homogeneous setting when all directions are fast, namely, when the population loss function is locally strongly convex, we defer the general theory of these settings to Appendix A. Here, we only consider the homogeneous settings where all directions are slow. To characterize the homogeneous settings, we assume that the population loss function f is locally convex in $\mathbb{B}(\theta^*, r)$ for some given radius r. Apart from the local convexity assumption, we also utilize the following assumption on the population loss function f.

(W.1) (Homogeneous Property) Given the constant $\alpha > 0$ and the radius $r > 0$, for all $\theta \in \mathbb{B}(\theta^*, r)$ we have

$$\lambda_{\min}(\nabla^2 f(\theta)) \geq c_1 \|\theta - \theta^*\|^{\alpha},$$
$$\lambda_{\max}(\nabla^2 f(\theta)) \leq c_2 \|\theta - \theta^*\|^{\alpha},$$

where $c_1 > 0$ and $c_2 > 0$ are some universal constants depending on r.

The condition $\alpha > 0$ is to ensure that the Hessian matrix is singular at the true parameter θ^*. For the setting $\alpha = 0$, corresponding to the locally strongly convex setting, the analysis of NormGD is in Appendix A. A simple example of Assumption (W.1) is $f(\theta) = \|\theta - \theta^*\|^{\alpha+2}$ for all $\theta \in \mathbb{B}(\theta^*, r)$. The Assumption (W.1) is satisfied by several statistical models, such as low signal-to-noise regime of generalized linear models with polynomial link functions (see Section 3.1) and symmetric two-component mixture model when the the true parameter is close to 0 (see Section 3.2). The homogeneous assumption (W.1) was also considered before to study the statistical and computational complexities of optimization algorithms [22].

Statistical rate of sample NormGD iterates θ_n^t: To establish the statistical and computational complexities of sample NormGD updates θ_n^t, we utilize the population to sample analysis [26, 1, 11, 13]. In particular, an application of triangle inequality leads to

$$\begin{aligned} \|\theta_n^{t+1} - \theta^*\| &\leq \|F_n(\theta_n^t) - F(\theta_n^t)\| + \|F(\theta_n^t) - \theta^*\| \\ &= A + B. \end{aligned} \tag{3}$$

Therefore, the statistical radius of θ_n^{t+1} around θ^* is controlled by two terms: (1) Term A: the uniform concentration of the sample NormGD operator F_n around the population GD operator F; (2) Term B: the contraction rate of population NormGD operator.

For term B in equation (3), the homogeneous assumption (W.1) entails the following contraction rate of population NormGD operator.

Lemma 1. *Assume Assumption (W.1) holds for some $\alpha > 0$ and some universal constants c_1, c_2. Then, if the step-size $\eta \leq \frac{c_1^2}{2c_2^2}$, then we have that*

$$\|F(\theta) - \theta^*\| \leq \kappa \|\theta - \theta^*\|,$$

where $\kappa < 1$ is a universal constant that only depends on η, c_1, c_2, α.

The proof of Lemma 1 is in Appendix B.1. For term A in equation (3), the uniform concentration bound between F_n and F, it can be obtained via the following assumption on the concentration bound of the operator norm of $\nabla^2 f_n(\theta) - \nabla^2 f(\theta)$ as long as $\|\theta - \theta^*\| \leq r$.

(W.2) (Stability of Second-order Information) For a given parameter $\gamma \geq 0$, there exist a noise function $\varepsilon : \mathbb{N} \times (0, 1] \to \mathbb{R}^+$, universal constant $c_3 > 0$, and some positive parameter $r > 0$ such that

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla^2 f_n(\theta) - \nabla^2 f(\theta)\|_{\text{op}} \leq c_3 r^{\gamma} \varepsilon(n, \delta),$$

for all $r \in (0, r)$ with probability $1 - \delta$.

To the best of our knowledge, the stability of second-order information in Assumption (W.2) is novel and has not been considered before to analyze the statistical guarantee of optimization algorithms. The idea of Assumption (W.2) is to control the growth of noise function, which is the difference between the population and sample loss functions, via the second-order information of these loss functions. A simple example for Assumption (W.2) is when $f_n(\theta) = \frac{\|\theta\|^{2p}}{2p} - \omega \frac{\|\theta\|^{2q}}{2q} \sqrt{\frac{d}{n}}$ where $\omega \sim \mathcal{N}(0, 1)$ and p, q are some positive integer numbers such that $p > q$. Then, $f(\theta) = \|\theta\|^{2p}/2p$. The Assumption (W.2) is satisfied with $\gamma = 2q - 2$ and with the noise function $\varepsilon(n, \delta) = \sqrt{\frac{d \log(1/\delta)}{n}}$. For concrete statistical examples, we demonstrate later in Section 3 that Assumption (W.2) is satisfied by generalized linear model and mixture model.

Given Assumption (W.2), we have the following uniform concentration bound between the sample NormGD operator F_n and population NormGD operator F.

Lemma 2. *Assume that Assumptions (W.1) and (W.2) hold with $\alpha \geq \gamma + 1$. Furthermore, assume that $\nabla f_n(\theta^*) = 0$. Then, we obtain that*

$$\sup_{\theta \in \mathbb{B}(\theta^*, r) \setminus \mathbb{B}(\theta^*, r_n)} \|F_n(\theta) - F(\theta)\| \leq c_4 r^{\gamma + 1 - \alpha} \varepsilon(n, \delta),$$

where $r_n := \left(\frac{6c_3 \epsilon(n,\delta)}{c_1}\right)^{\frac{1}{\alpha - \gamma}}$, and c_4 is a universal constant depends on $\eta, c_1, c_2, c_3, \alpha, \gamma$.

The proof of Lemma 2 is in Appendix B.2. We have a few remarks with Lemma 2. First, the assumption that $\nabla f_n(\theta^*) = 0$ is to guarantee the stability of $\nabla f_n(\theta)$ around $\nabla f(\theta)$ as long as $\|\theta - \theta^*\| \leq r$ for any $r > 0$ [11]. Second, the assumption that $\alpha \geq \gamma + 1$ means that the signal is stronger than the noise in statistical models, which in turn leads to meaningful statistical rates. Third, the inner radius r_n in Lemma 2 corresponds to the final statistical radius, which is at the order $\mathcal{O}(\epsilon(n,\delta)^{\frac{1}{\alpha-\gamma}})$. It means that we cannot go beyond that radius, or otherwise the empirical Hessian is not positive definite.

Based on the contraction rate of population NormGD operator in Lemma 1 and the uniform concentration of the sample NormGD operator around the population NormGD operator in Lemma 2, we have the following result on the statistical and computational complexities of the sample NormGD iterates around the true parameter θ^*.

Theorem 1. *Assume that Assumptions (W.1) and (W.2) and assumptions in Lemma 2 hold with $\alpha \geq \gamma+1$. Assume that the sample size n is large enough such that $\varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}} \leq \frac{(1-\kappa)r}{c_4 \bar{C}^{\gamma+1-\alpha}}$ where κ is defined in Lemma 1, c_4 is the universal constant in Lemma 2 and $\bar{C} = (\frac{6c_3}{c_1})^{\frac{1}{\alpha-\gamma}}$, and r is the local radius. Then, there exist universal constants C_1, C_2 such that with probability $1-\delta$, for $t \geq C_1 \log(1/\varepsilon(n,\delta))$, the following holds:*

$$\min_{k\in\{0,1,\cdots,t\}} \|\theta_n^k - \theta^*\| \leq C_2 \cdot \varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}}.$$

The proof of Theorem 1 is in Appendix B.3. A few comments with Theorem 1 are in order.

On the approximation of $\lambda_{\max}$: Computing the whole spectrum of a $d \times d$ matrix requires $\mathcal{O}(d^3)$ computation. But fortunately, we can compute the maximum eigenvalue in $\mathcal{O}(d^2)$ computation with the well-known power iteration [*a.k.a* power method, see Chapter 7.3, 9]) which has broad applications in different areas [e.g. 10]. Power iteration can compute the maximum eigenvalue up to ε error with at most $\mathcal{O}\left(\frac{\log \varepsilon}{\log(\lambda_2/\lambda_{\max})}\right)$ matrix vector products, where λ_2 is the second largest eigenvalue. Hence, when $\lambda_2/\lambda_{\max}$ is bounded away from 1, we can obtain a high-quality approximation of $\lambda_{\max}$ with small number of computation. Things can be a little weird when $\lambda_2/\lambda_{\max}$ is close to 1. But in fact, we only requires an approximation of $\lambda_{\max}$ within statistical accuracy defined in Assumption (W.2). Hence, without loss of generality, we can assume $\lambda_2(\nabla^2 f_n(\theta)) \leq \lambda_{\max}(\nabla^2 f_n(\theta)) - c_3\|\theta-\theta^*\|^{\gamma}\varepsilon(n,\delta)$, which means $\frac{\lambda_2(\nabla^2 f_n(\theta))}{\lambda_{\max}(\nabla^2 f_n(\theta))} \leq 1 - \frac{c_3}{c_2}\|\theta-\theta^*\|^{\gamma-\alpha}$. Since $\alpha \geq \gamma+1$ and we only consider the case $\|\theta-\theta^*\| \leq r$, we know their exists a universal constant $c_{\mathrm{PI}} < 1$ that does not depend on n, d, such that $\lambda_2/\lambda_{\max} \leq c_{\mathrm{PI}}$. As a result, we can always compute the $\lambda_{\max}$ with small number of iterations.

Comparing to fixed-step size gradient descent: Under the Assumptions (W.1) and (W.2), we have the following result regarding the statistical and computational complexities of fixed-step size GD iterates.

Proposition 1. *Assume that Assumptions (W.1) and (W.2) hold with $\alpha \geq \gamma+1$ and $\nabla f_n(\theta^*) = 0$. Suppose the sample size n is large enough so that $\varepsilon(n,\delta) \leq C$ for some universal constant C. Then there exist universal constant C_1 and C_2, such that for any fixed $\tau \in \left(0, \frac{1}{\alpha-\gamma}\right)$, as long as $t \geq C_1 \varepsilon(n,\delta)^{-\frac{\alpha}{\alpha-\gamma}} \log\frac{1}{\tau}$, we have that*

$$\|\theta_{n,GD}^t - \theta^*\| \leq C_2 \varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}-\tau}.$$

The proof of Proposition 1 is in Appendix B.4. Therefore, the results in Theorem 1 indicate that the NormGD and fixed-step size GD iterates reach the same statistical radius $\varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}}$ within the true parameter θ^*. Nevertheless, the NormGD only takes $\mathcal{O}(\log(1/\varepsilon(n,\delta)))$ number of iterations while the fixed-step size GD takes $\mathcal{O}(\varepsilon(n,\delta)^{-\frac{\alpha}{\alpha-\gamma}})$ number of iterations. If the dimension d is fixed, the total computational complexity of NormGD algorithm is at the order of $\mathcal{O}(n \cdot \log(1/\varepsilon(n,\delta)))$, which is much cheaper than that of fixed-step size GD, $\mathcal{O}(n \cdot \varepsilon(n,\delta)^{-\frac{\alpha}{\alpha-\gamma}})$, to reach the final statistical radius.

3 Examples

In this section, we consider an application of our theories in previous section to the generalized linear model and Gaussian mixture model.

3.1 Generalized Linear Model (GLM)

Generalized linear model (GLM) has been a widely used model in statistics and machine learning [18]. It is a generalization of linear regression model where we use a link function to relate the covariates to the response variable. In particular, we assume that $(Y_1, X_1), \ldots, (Y_n, X_n) \in \mathbb{R} \times \mathbb{R}^d$ satisfy

$$Y_i = g(X_i^\top \theta^*) + \varepsilon_i. \qquad \forall i \in [n] \tag{4}$$

Here, $g : \mathbb{R} \to \mathbb{R}$ is a given link function, θ^* is a true but unknown parameter, and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. noises from $\mathcal{N}(0, \sigma^2)$ where $\sigma > 0$ is a given variance parameter. We consider the random design setting where $X_1, \ldots, X_n$ are i.i.d. from $\mathcal{N}(0, I_d)$. A few comments with our model assumption. First, in our paper, we will not estimate the link function g. Second, the assumption that the noise follows the Gaussian distribution is just for the simplicity of calculations; similar proof argument still holds for sub-Gaussian noise. For the purpose of our theory, we consider the link function $g(r) := r^p$ for any $p \in \mathbb{N}$ and $p \geq 2$. When $p = 2$, the generalize linear model becomes the phase retrieval problem [8, 24, 4, 20].

Least-square loss: We estimate the true parameter θ^* via minimizing the least-square loss function, which is:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}_n(\theta) := \frac{1}{2n} \sum_{i=1}^{n} (Y_i - (X_i^\top \theta)^p)^2. \tag{5}$$

By letting the sample size n goes to infinity, we obtain the population least-square loss function of GLM:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) := \frac{1}{2} \mathbb{E}_{X,Y}[(Y - (X^\top \theta)^p)^2],$$

where the outer expectation is taken with respect to $X \sim \mathcal{N}(0, I_d)$ and $Y = g(X^\top \theta^*) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. It is clear that θ^* is the global minimum of the population loss function $\mathcal{L}$. Furthermore, the function $\mathcal{L}$ is homogeneous, i.e., all directions have similar behaviors.

In this section, we consider two regimes of the GLM for our study of sample NormGD iterates: Strong signal-to-noise regime and Low signal-to-noise regime.

Strong signal-to-noise regime: The strong signal-to-noise regime corresponds to the setting when θ^* is bounded away from 0 and $\|\theta^*\|$ is sufficiently large, i.e., $\|\theta^*\| \geq C$ for some universal constant C. Under this setting, we can check that the population loss function $\mathcal{L}$ is locally strongly convex and smooth, i.e., it satisfies Assumption (S.1) under the homogeneous setting with all fast directions. Furthermore, for Assumption (S.2), for any radius $r > 0$ there exist universal constants C_1, C_2, C_3 such that as long as $n \geq C_1(d \log(d/\delta))^{2p}$, the following uniform concentration bounds hold:

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla \mathcal{L}_n(\theta) - \nabla \mathcal{L}(\theta)\| \leq C_2 \sqrt{\frac{d + \log(1/\delta)}{n}}, \tag{6}$$

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla^2 \mathcal{L}_n(\theta) - \nabla^2 \mathcal{L}(\theta)\|_{\text{op}} \leq C_3 \sqrt{\frac{d + \log(1/\delta)}{n}}. \tag{7}$$

The proof can be found in Appendix C.2.

Low signal-to-noise regime: The low signal-to-noise regime corresponds to the setting when the value of $\|\theta^*\|$ is sufficiently small. To simplify the computation, we assume that $\theta^* = 0$. Direct calculation shows that the population loss function becomes

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) = \frac{\sigma^2 + (2p-1)!!\|\theta - \theta^*\|^{2p}}{2}. \tag{8}$$

Under this setting, the population loss function $\mathcal{L}$ is no longer locally strong convex around $\theta^* = 0$. Indeed, this function is homogeneous with all slow directions, which are given by:

$$\lambda_{\max}(\nabla^2 \mathcal{L}(\theta)) \leq c_1 \|\theta - \theta^*\|^{2p-2}, \tag{9}$$

$$\lambda_{\min}(\nabla^2 \mathcal{L}(\theta)) \geq c_2 \|\theta - \theta^*\|^{2p-2}, \tag{10}$$

for all $\theta \in \mathbb{B}(\theta^*, r)$ for some $r > 0$. Here, c_1, c_2 are some universal constants depending on r. Therefore, the homogeneous Assumption (W.1) is satisfied with $\alpha = 2p - 2$. The proof for the claims (16) and (17) is in Appendix C.1.

Moving to Assumption (W.2), we demonstrate in Appendix C.2 that we can find universal constants C_1 and C_2 such that for any $r > 0$ and $n \geq C_1(d \log(d/\delta))^{2p}$ we have

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla^2 \mathcal{L}_n(\theta) - \nabla^2 \mathcal{L}(\theta)\|_{\text{op}} \leq C_2 (r^{p-2} + r^{2p-2}) \sqrt{\frac{d + \log(1/\delta)}{n}} \tag{11}$$

with probability at least $1 - \delta$. Hence, the stability of second order information Assumption (W.2) is satisfied with $\gamma = p - 2$.

Based on the above results, Theorems 1 for homogeneous settings with all slow directions and 2 for homogeneous settings with all fast directions lead to the following statistical and computational complexities of NormGD algorithm for solving the true parameter of GLM.

Corollary 1. *Given the generalized linear model (4) with $g(r) = r^p$ for some $p \in \mathbb{N}$ and $p \geq 2$, there exists universal constants $c, \tilde{c}_1, \tilde{c}_2, \bar{c}_1, \bar{c}_2$ such that when the sample size $n \geq c(d \log(d/\delta))^{2p}$ and the initialization $\theta_n^0 \in \mathbb{B}(\theta^*, r)$ for some chosen radius $r > 0$, with probability $1 - \delta$ the sequence of sample NormGD iterates $\{\theta_n^t\}_{t \geq 0}$ satisfies the following bounds:*

(i) Strong signal-to-noise regime: When $\|\theta^*\| \geq C$ for some universal constant C, we find that

$$\|\theta_n^t - \theta^*\| \leq \tilde{c}_1 \sqrt{\frac{d + \log(1/\delta)}{n}}, \qquad \textit{for } t \geq \tilde{c}_2 \log\left(\frac{n}{d + \log(1/\delta)}\right),$$

(ii) Low signal-to-noise regime: When $\theta^* = 0$, we obtain

$$\min_{1 \leq k \leq t} \|\theta_n^k - \theta^*\| \leq c_1' \left(\frac{d + \log(1/\delta)}{n}\right)^{1/(2p)}, \qquad \textit{for } t \geq c_2' \log\left(\frac{n}{d + \log(1/\delta)}\right).$$

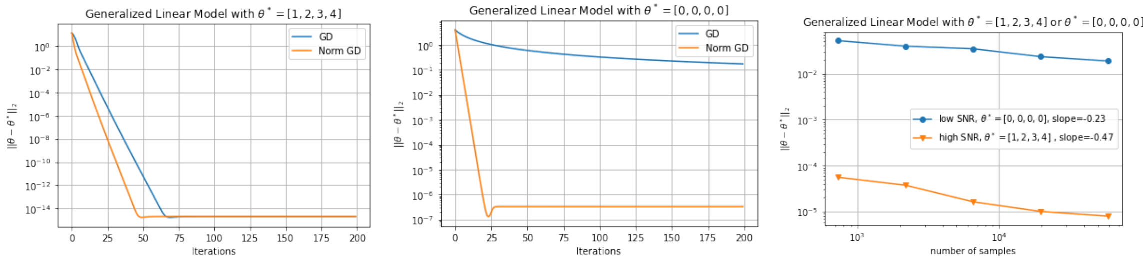

Figure 1. *Verification simulation for the Generalized Linear Model (GLM) example.* **Left**: Both GD and Norm GD converges linearly in the high signal-to-noise setting; **Middle**: only Norm GD converges linearly in the low signal-to-noise setting while GD converges sub-linearly; **Right**: the log-log plot of sample size versus statistical error shows that the statistical error scales with $n^{-0.5}$ in the strong signal-to-noise setting and $n^{-0.25}$ in the low signal-to-noise setting, which coincides with our theory. The slope is computed as the linear regression coefficient of the log sample size versus the log statistical error.

A few comments with Corollary 1 are in order. For the strong signal-to-noise regime, the sample NormGD only takes logarithmic number of iterations $\log(n)$ to reach the optimal statistical radius $(d/n)^{1/2}$ around the true parameter. This guarantee is similar to that of the fixed-step size GD iterates for solving the locally strongly convex and smooth loss function [1, 11]. For the low signal-to-noise regime, the sample NormGD iterates reach the final statistical radius $(d/n)^{1/2p}$ after logarithmic number of iterations in terms of n. In terms of the number of iterations, it is cheaper than that of that fixed-step size GD algorithm, which takes at least $\mathcal{O}(n^{\frac{p-1}{p}})$ number of iterations for fixed dimension d (See our discussion after Theorem 1). For fixed d, it indicates that the total computational complexity of NormGD algorithm, which is at the order of $\mathcal{O}(n)$, is smaller than that of fixed-step size GD, which is $\mathcal{O}(n^{1+\frac{p-1}{p}})$. Therefore, for the low signal-to-noise regime, the NormGD algorithm is more computationally efficient than the fixed-step size GD algorithm for reaching the similar final statistical radius.

Experiments: To verify our theory, we performed simulation on generalized linear model, and the results are shown in Figure 1. We set $p = 2$ and $d = 4$. For the low signal-to-noise setting, we set θ^* to be $[0, 0, 0, 0]$, and for high signal-to-noise setting, we set θ^* to be $[1, 2, 3, 4]$. For the left and the middle plots in Figure 1, the sample size is set to be 1000. As in the left plot, when in the strong signal-to-noise setting, both the fixed step size Gradient Descent method (referred to as GD) and our proposed Normalized Gradient Descent method (referred to as NormGD) converges linearly. However, once we shift to the low signal-to-noise setting, only Norm GD converges linearly, while GD converges only sub-linearly, as shown in the middle plot of Figure 1. To further verify our corollaries, especially how the statistical error scales with n, we plot the statistical error versus sample size in the right plot as in Figure 1. The experiments was repeated for 10 times and the average of the statistical error is shown. The slope is computed as the linear regression coefficient of the log sample size versus the log statistical error. As in this log-log plot, in the strong signal-to-noise setting, the statistical error roughly scales with $n^{-0.5}$, while in the low signal-to-noise setting, the statistical error roughly scales with $n^{-0.25}$. This coincides with our theory as in Corollary 1.

3.2 Gaussian mixture models (GMM)

We now consider Gaussian mixture models (GMM), one of the most popular statistical models for modeling heterogeneous data [14, 15]. Parameter estimation in these models plays an important role in capturing the heterogeneity of different subpopulations. The common approach to estimate the location and scale parameters in these model is via maximizing the log-likelihood function. The statistical guarantee of the maximum likelihood estimator (MLE) in Gaussian mixtures had been studied in [5, 12]. However, since the log-likelihood function is highly non-concave, in general we do not have closed-form expressions for the MLE. Therefore, in practice we utilize optimization algorithms to approximate the MLE. However, a complete picture about the statistical and computational complexities of these optimization algorithms have remained poorly understood.

In order to shed light on the behavior of NormGD algorithm for solving GMM, we consider a simplified yet important setting of this model, symmetric two-component location GMM. This model had been used in the literature to study the statistical behaviors of Expectation-Maximization (EM) algorithm [1, 7]. We assume that the data $X_1, X_2, \ldots, X_n$ are i.i.d. samples from $\frac{1}{2}\mathcal{N}(-\theta^*, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\theta^*, \sigma^2 I_d)$ where $\sigma > 0$ is given and θ^* is true but unknown parameter. Our goal is to obtain an estimation of θ^* via also using the symmetric two-component location Gaussian mixture:

$$\frac{1}{2}\mathcal{N}(-\theta, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\theta, \sigma^2 I_d). \tag{12}$$

As we mentioned earlier, we obtain an estimation of θ^* via maximizing the sample log-likelihood function associated with model (12), which admits the following form:

$$\min_{\theta \in \mathbb{R}^d} \bar{\mathcal{L}}_n(\theta) := -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{2}\phi(X_i|\theta, \sigma^2 I_d) + \frac{1}{2}\phi(X_i| - \theta, \sigma^2 I_d)\right). \tag{13}$$

Here, $\phi(\cdot|\theta, \sigma^2 I_d)$ denotes the density function of multivariate Gaussian distribution with mean θ and covariance matrix $\sigma^2 I_d$.

Similar to GLM, we also consider two regimes of the true parameter: Strong signal-to-noise regime when $\|\theta^*\|/\sigma$ is sufficiently large and Low signal-to-noise regime when $\|\theta^*\|/\sigma$ is sufficiently small. To analyze the behaviors of sample NormGD iterates, we define the population version of the maximum likelihood estimation (13) as follows:

$$\min_{\theta \in \mathbb{R}^d} \bar{\mathcal{L}}(\theta) := -\mathbb{E}\left[\log\left(\frac{1}{2}\phi(X|\theta, \sigma^2 I_d) + \frac{1}{2}\phi(X| - \theta, \sigma^2 I_d)\right)\right]. \tag{14}$$

Here, the outer expectation is taken with respect to $X \sim \frac{1}{2}\mathcal{N}(-\theta^*, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(\theta^*, \sigma^2 I_d)$. We can check that $\bar{\mathcal{L}}$ is also homogeneous in all directions. The strong signal-to-noise regime corresponds to the setting when $\bar{\mathcal{L}}$ is homogeneous with all fast directions while the low signal-to-noise regime is associated with the setting when $\bar{\mathcal{L}}$ is homogeneous with all slow directions.

Strong signal-to-noise regime: For the strong signal-to-noise regime, we assume that $\|\theta^*\| \geq C\sigma$ for some universal constant C. Since the function $\bar{\mathcal{L}}$ is locally strongly convex and smooth as long as $\theta \in \mathbb{B}(\theta^*, \frac{\|\theta^*\|}{4})$ (see Corollary 1 in [1]), the Assumption (S.1) under the homogeneous setting with all fast directions is satisfied. Furthermore, as long as we choose the radius $r \leq \frac{\|\theta^*\|}{4}$ and the sample size $n \geq C_1 d \log(1/\delta)$ for some universal constant C_1, with probability at least $1 - \delta$ there exist universal constants C_2 and C_3 such that

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla \bar{\mathcal{L}}_n(\theta) - \nabla \bar{\mathcal{L}}(\theta)\| \leq C_2 \sqrt{\frac{d \log(1/\delta)}{n}},$$

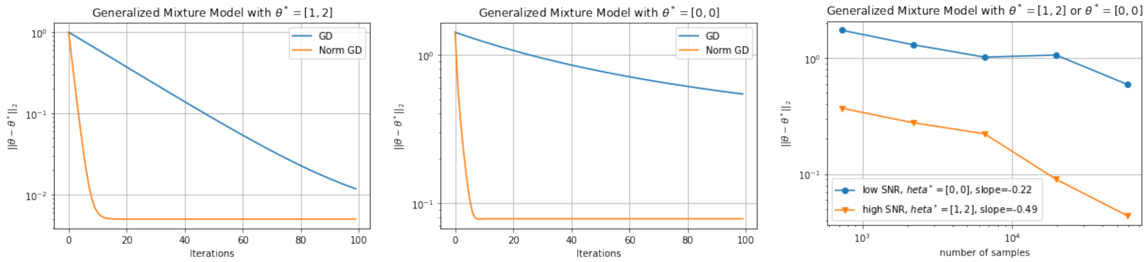

Figure 2. *Verification simulation for the Gaussian Mixture Model (GMM) example.* **Left**: Both GD and Norm GD converges linearly in the strong signal-to-noise setting; **Middle**: only Norm GD converges linearly in the low signal-to-noise setting while GD converges sub-linearly; **Right**: the log-log plot of sample size versus statistical error shows that the statistical error scales with $n^{-0.5}$ in the strong signal-to-noise setting and $n^{-0.25}$ in the low signal-to-noise setting, which coincides with our theory. The slope is computed as the linear regression coefficient of the log sample size versus the log statistical error.

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla^2 \bar{\mathcal{L}}_n(\theta) - \nabla^2 \bar{\mathcal{L}}(\theta)\|_{\text{op}} \leq C_3 \sqrt{\frac{d \log(1/\delta)}{n}}. \tag{15}$$

The proof of claims (15) is in Appendix D.2. In light of Theorem 2 in Appendix A for homogeneous settings with all fast directions, the NormGD iterates converge to the final statistical radius $(d/n)^{1/2}$ after $\log(n)$ iterations (see Corollary 2 for a formal statement of this result).

Low signal-to-noise regime: Moving to the low signal-to-noise regime, which refers to the setting when $\|\theta^*\|/\sigma$ is sufficiently small. For the simplicity of computation we specifically assume that $\theta^* = 0$. Under this setting, the true model becomes a single Gaussian distribution with mean 0 and covariance matrix $\sigma^2 I_d$ while the fitted model (12) has two components with similar weights and symmetric means. This setting is widely referred to as over-specified mixture model, namely, we fit the true mixture model with more components than needed, in statistics and machine learning [5, 23]. It is important in practice as the true number of components is rarely known and to avoid underfitting the true model, we tend to use a fitted model with more components than the true number of components.

In Appendix D.1, we prove that the population loss function $\bar{\mathcal{L}}$ is homogeneous with all slow directions and satisfy the following properties:

$$\lambda_{\max}(\nabla^2 \bar{\mathcal{L}}(\theta)) \leq c_1 \|\theta - \theta^*\|^2, \tag{16}$$

$$\lambda_{\min}(\nabla^2 \bar{\mathcal{L}}(\theta)) \geq c_2 \|\theta - \theta^*\|^2, \tag{17}$$

for all $\theta \in \mathbb{B}(\theta^*, \frac{\sigma}{2})$ where c_1 and c_2 are some universal constants. Therefore, the population loss function $\bar{\mathcal{L}}$ satisfies Assumption (W.1) with $\alpha = 2$.

For the stability of second-order information, we prove in Appendix D.2 that there exist universal constants C_1 and C_2 such that for any $r > 0$, with probability $1 - \delta$ as long as $n \geq C_1 d \log(1/\delta)$ we obtain

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla^2 \bar{\mathcal{L}}_n(\theta) - \nabla^2 \bar{\mathcal{L}}(\theta)\| \leq C_2 \sqrt{\frac{d \log(1/\delta)}{n}}. \tag{18}$$

The uniform concentration bound (18) shows that for the low signal-to-noise regime of two-component location Gaussian mixtures, the stability of second-order information in Assumption (W.2) is satisfied with $\gamma = 0$. Moreover, from Lemma 1 in [6] we know $\nabla f_n(\theta) = 0$.

Combining the results from the homogeneous behaviors of population loss function in equations (16)-(17) and the uniform concentration bound in equation (18) to the result of Theorem 1, we obtain that the NormGD updates reach the final statistical radius $(d/n)^{1/4}$ after $\log(n)$ number of iterations.

Now, we would like to formally state the statistical behaviors of the NormGD iterates for both the strong signal-to-noise and low signal-to-noise regimes.

Corollary 2. *Given the symmetric two-component mixture model (12), we can find positive universal constants $c, \bar{c}_1, \bar{c}_2, c_1', c_2'$ such that with probability at least $1-\delta$, when $n \geq cd\log(1/\delta)$ the sequence of NormGD iterates $\{\theta_n^t\}_{t\geq 0}$ satisfies the following bounds:*

(i) Strong signal-to-noise regime: When $\|\theta^*\| \geq C$ for some sufficiently large constant C and the initialization $\theta_n^0 \in \mathbb{B}(\theta^*, \frac{\|\theta\|^*}{4})$, we obtain that

$$\|\theta_n^t - \theta^*\| \leq \bar{c}_1 \sqrt{\frac{d\log(1/\delta)}{n}}, \qquad \textit{as long as } t \geq \bar{c}_2 \log\left(\frac{n}{d\log(1/\delta)}\right),$$

(ii) Low signal-to-noise regime: Under the setting $\theta^* = 0$ and the initialization $\theta_n^0 \in \mathbb{B}(\theta^*, \frac{\sigma}{2})$, we have

$$\min_{1\leq k\leq t} \|\theta_n^k - \theta^*\| \leq c_1' \left(\frac{d\log(1/\delta)}{n}\right)^{1/4}, \qquad \textit{for } t \geq c_2' \log\left(\frac{n}{d\log(1/\delta)}\right).$$

We have the following comments with the results of Corollary 2. In the strong signal-to-noise case, the NormGD algorithm and the fixed step size GD algorithm, which is also the EM algorithm for the symmetric two-component mixture, reach the final statistical radius $(d/n)^{1/2}$ around the true parameter θ^* after $\log(n)$ number of iterations. For the low signal-to-noise regime, the NormGD iterates reach the final statistical radius $(d/n)^{1/4}$ after a logarithmic number of iterations in terms of n while the EM iterates reach that radius after $\sqrt{n}$ number of iterations [7]. It demonstrates that for fixed dimension d the total computational complexity for the NormGD is at the order of $\mathcal{O}(n)$, which is much cheaper than that of the EM algorithm, which is at the order of $\mathcal{O}(n^{3/2})$.

Experiments: To verify our theory, we performed simulation on Gaussian Mixture Model (GMM), and the results are shown in Figure 2. We set $d = 2$. For the low signal-to-noise setting, we set θ^* to be $[0, 0]$, and for strong signal-to-noise setting, we set θ^* to be $[1, 2]$. For the left and the middle plots in Figure 1, the sample size is set to be 10000. As in the left plot, when in the strong signal-to-noise setting, both the fixed step size Gradient Descent method (referred to as GD, and is essentially EM algorithm as described above) and our proposed Normalized Gradient Descent method (referred to as NormGD) converges linearly. However, once we shift from the strong signal-to-noise setting to the low signal-to-noise setting, only Norm GD converges linearly, while GD converges only sub-linearly, as shown in the middle plot. To further verify our corollaries, especially how the statistical error scales with n, we plot the statistical error versus sample size in the right plot. The experiments were repeated for 10 times and the average of the statistical error is shown. The slope is computed as the linear regression coefficient of the log sample size versus the log statistical error. As in this log-log plot, in the strong signal-to-noise setting, the statistical error roughly scales with $n^{-0.5}$, while in the low signal-to-noise setting, the statistical error roughly scales with $n^{-0.25}$. This coincides with our theory as in Corollary 2.

4 Conclusion

In this paper, we show that by utilizing second-order information in the design of optimization algorithms, we are able to improve the computational complexity of these algorithms for solving parameter estimation in statistical models. In particular, we study the statistical and computational complexities of the NormGD algorithm, a variant of gradient descent algorithm whose step size is scaled by the maximum eigenvalue of the Hessian matrix of the loss function. We show that when the population loss function is homogeneous, the NormGD algorithm only needs a logarithmic number of iterations to reach the final statistical radius around the true parameter. In terms of iteration complexity and total computational complexity, it is cheaper than fixed step size GD algorithm, which requires a polynomial number of iterations to reach the similar statistical radius under the singular statistical model settings.

5 Acknowledgements

This work was partially supported by the NSF IFML 2019844 award and research gifts by UT Austin ML grant to NH, and by NSF awards 1564000 and 1934932 to SS.

References

[1] S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45:77–120, 2017. (Cited on pages 2, 3, 4, 9, 10, 16, and 34.)

[2] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013. (Cited on pages 25, 28, and 31.)

[3] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. (Cited on pages 1 and 21.)

[4] E. J. Candes, Y. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion, 2011. (Cited on page 7.)

[5] J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233, 1995. (Cited on pages 10 and 11.)

[6] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Sharp analysis of expectation-maximization for weakly identifiable models. *AISTATS*, 2020. (Cited on pages 4 and 12.)

[7] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Singularity, misspecification, and the convergence rate of EM. *Annals of Statistics*, 44:2726–2755, 2020. (Cited on pages 3, 10, and 12.)

[8] J. R. Fienup. Phase retrieval algorithms: a comparison. *Appl. Opt.*, 21(15):2758–2769, Aug 1982. (Cited on page 7.)

[9] G. H. Golub and C. F. Van Loan. Matrix computations. johns hopkins studies in the mathematical sciences, 1996. (Cited on page 6.)

[10] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. (Cited on page 6.)

[11] N. Ho, K. Khamaru, R. Dwivedi, M. J. Wainwright, M. I. Jordan, and B. Yu. Instability, computational efficiency and statistical accuracy. *Arxiv Preprint Arxiv: 2005.11411*, 2020. (Cited on pages 2, 4, 5, 9, 16, and 22.)

[12] N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016. (Cited on page 10.)

[13] J. Y. Kwon, N. Ho, and C. Caramanis. On the minimax optimality of the EM algorithm for learning two-component mixed linear regression. In *AISTATS*, 2021. (Cited on page 4.)

[14] B. Lindsay. *Mixture Models: Theory, Geometry and Applications.* In NSF-CBMS Regional Conference Series in Probability and Statistics. IMS, Hayward, CA., 1995. (Cited on page 9.)

[15] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs.* New York, 1988. (Cited on page 9.)

[16] J. Mei, Y. Gao, B. Dai, C. Szepesvari, and D. Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. *arXiv preprint arXiv:2105.06072*, 2021. (Cited on page 21.)

[17] W. Mou, N. Ho, M. J. Wainwright, P. Bartlett, and M. I. Jordan. A diffusion process perspective on posterior contraction rates for parameters. *arXiv preprint arXiv:1909.00966*, 2019. (Cited on pages 25, 28, 30, and 31.)

[18] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135, 1972. (Cited on page 7.)

[19] Y. Nesterov. *Lectures on Convex Optimization.* Springer, 2018. (Cited on page 1.)

[20] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. *IEEE Transactions on Signal Processing*, 63(18):4814–4826, 2015. (Cited on page 7.)

[21] B. T. Polyak. *Introduction to Optimization.* Optimization Software, Inc., New York, 1987. (Cited on page 1.)

[22] T. Ren, F. Cui, A. Atsidakou, S. Sanghavi, and N. Ho. Towards statistical and computational complexities of Polyak step size gradient descent. In *AISTATS, 2022*, 2022. (Cited on pages 3, 4, and 23.)

[23] J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:689–710, 2011. (Cited on page 11.)

[24] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015. (Cited on page 7.)

[25] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. (Cited on pages 24 and 34.)

[26] X. Yi and C. Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015. (Cited on page 4.)

[27] J. Zhuo, J. Kwon, N. Ho, and C. Caramanis. On the computational and statistical complexity of over-parameterized matrix sensing. *arViv preprint arXiv: 2102.02756*, 2021. (Cited on page 4.)

Supplement to "Improving Computational Complexity in Statistical Models with Second-Order Information"

In the supplementary material, we collect proofs and results deferred from the main text. In Appendix A, we provide general theory for the statistical guarantee of NormGD for the homogeneous settings with all fast directions of the population loss function. In Appendix B, we provide proofs for the main results in the main text. We then provide proofs for the statistical and computational complexities of NormGD under generalized linear models and mixture models respectively in Appendices C and D.

A Homogeneous Settings with All Fast Directions

In this Appendix, we provide statistical guarantee for the NormGD iterates when the population loss function is homogeneous with all fast directions. Following the population to sample analysis in equation (3), we first consider the strong convexity and Lipschitz smoothness assumptions that characterize all fast directions.

(S.1) (Strongly convexity and Lipschitz smoothness) For some radius $r > 0$, for all $\theta \in \mathbb{B}(\theta^*, r)$ we have

$$\bar{c}_1 \leq \lambda_{\min}(\nabla^2 f(\theta)) \leq \lambda_{\max}(\nabla^2 f(\theta)) \leq \bar{c}_2,$$

where $\bar{c}_1 > 0$ and $\bar{c}_2 > 0$ are some universal constants depending on r.

The Assumption (S.1) is a special case of Assumption (W.1) when $\alpha = 0$. A simple example for the function f that satisfies Assumption (S.1) is $f(\theta) = \|\theta\|^2$.

Given the Assumption (S.1), we obtain the following result for the contraction of the population NormGD operator F around the true parameter θ^*.

Lemma 3. *Assume Assumption (S.1) holds for some universal constants $\bar{c}_1, \bar{c}_2$. Then, if the step-size $\eta \leq \frac{\bar{c}_2^2}{2\bar{c}_1^2}$, then we have that*

$$\|F(\theta)) - \theta^*\| \leq \bar{\kappa}\|\theta - \theta^*\|,$$

where $\bar{\kappa} < 1$ is a universal constant that only depends on $\eta, \bar{c}_1, \bar{c}_2$.

The proof of Lemma 3 is a direct from the proof of Lemma 1 with $\alpha = 0$; therefore, its proof is omitted.

(S.2) (Stability of first and second-order information) For some fixed positive parameter $r > 0$, there exist a noise function $\varepsilon : \mathbb{N} \times (0, 1] \to \mathbb{R}^+$, and universal constants $\bar{c}_3, \bar{c}_4 > 0$ depends on r, such that

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla f_n(\theta) - \nabla f(\theta)\| \leq \bar{c}_3 \cdot \varepsilon(n, \delta),$$
$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla^2 f_n(\theta) - \nabla^2 f(\theta)\|_{\text{op}} \leq \bar{c}_4 \cdot \varepsilon(n, \delta).$$

for all $r \in (0, r)$ with probability $1 - \delta$.

We would like to remark that the assumption in the uniform concentration of $\nabla f_n(\theta)$ around $\nabla f(\theta)$ is standard for analyzing optimization algorithms for solving parameter estimation under locally strongly convex and smooth population loss function [1, 11]. The extra assumption on the uniform concentration of the empirical Hessian matrix $\nabla^2 f_n(\theta)$ around the population Hessian matrix $\nabla^2 f(\theta)$ is to ensure that $\lambda_{\max}(\nabla^2 f_n(\theta)$ in NormGD algorithm will stay close to $\lambda_{\max}(\nabla^2 f(\theta))$. These two conditions are sufficient to guarantee the stability of the sample NormGD operator F_n around the population NormGD operator in the following lemma.

Lemma 4. *Assume that Assumption (S.2) holds, and n is sufficiently large such that $\bar{c}_1 > 2\bar{c}_3\varepsilon(n, \delta)$. Then, we obtain that*

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|F_n(\theta) - F(\theta)\| \leq \bar{c}_5 \varepsilon(n, \delta),$$

and $\bar{c}_5$ is a universal constant depends on $\eta, \bar{c}_1, \bar{c}_2, \bar{c}_3, \bar{c}_4$.

Proof. With straightforward calculation, we have that

$$\begin{aligned}
\|F_n(\theta) - F(\theta)\| \leq & \eta \left(\left\| \frac{\nabla f(\theta)(\lambda_{\max}(\nabla^2 f(\theta)) - \lambda_{\max}(\nabla^2 f_n(\theta)))}{\lambda_{\max}(\nabla^2 f_n(\theta))\lambda_{\max}(\nabla^2 f(\theta))} \right\| + \left\| \frac{\nabla f_n(\theta) - \nabla f(\theta)}{\lambda_{\max}(\nabla^2 f_n(\theta))} \right\| \right) \\
\leq & \eta \left(\frac{\bar{c}_2 \bar{c}_3 \epsilon(n, \delta)}{(\bar{c}_1 - \bar{c}_3 \epsilon(n, \delta))\bar{c}_1} + \frac{\bar{c}_4 \epsilon(n, \delta)}{\bar{c}_1 - \bar{c}_3 \epsilon(n, \delta)} \right) \\
\leq & \eta \left(\frac{2\bar{c}_2\bar{c}_3 + 2\bar{c}_1\bar{c}_4}{\bar{c_1}^2} \right) \varepsilon(n, \delta).
\end{aligned}$$

Take $\bar{c}_5$ accordingly, we conclude the proof. $\square$

Theorem 2. *Assume Assumptions (S.1) and (S.2) hold, and n is sufficient large such that $\bar{c}_1 > 2\bar{c}_3\varepsilon(n, \delta)$ and $\bar{c}_5\varepsilon(n, \delta) \leq (1 - \bar{\kappa})r$ where $\bar{\kappa}$ is the constant defined in Lemma 3. Then, there exist universal constants $\bar{C}_1$, $\bar{C}_2$ such that for $t \geq \bar{C}_1 \log(1/\varepsilon(n, \delta))$, the following holds:*

$$\|\theta_n^t - \theta^*\| \leq \bar{C}_2 \cdot \varepsilon(n, \delta),$$

Proof. With the triangle inequality, we have that

$$\begin{aligned}
\|\theta_n^{k+1} - \theta^*\| = & \|F_n(\theta_n^k) - \theta^*\| \\
\leq & \|F_n(\theta_n^k) - F(\theta_n^k)\| + \|F(\theta_n^k) - \theta^*\| \\
\leq & \sup_{\theta \in \mathbb{B}(\theta^*, r)} \|F_n(\theta) - F(\theta)\| + \kappa\|\theta_n^k - \theta^*\|
\end{aligned}$$

$$\le \bar{c}_5 \varepsilon(n,\delta) + \bar{\kappa} r$$
$$\le r.$$

Hence, we know $\|\theta_n^t - \theta^*\| \le r, \forall t \in \mathbb{N}$. Furthermore, by repeating the above argument T times, we can obtain that

$$\begin{aligned}\|\theta_n^T - \theta^*\| \le& \bar{c}_5 \varepsilon(n,\delta) \left(\sum_{t=0}^{T-1} \bar{\kappa}^t\right) + \kappa^T \|\theta_n^0 - \theta^*\| \\ \le& \frac{\bar{c}_5}{1-\bar{\kappa}} \varepsilon(n,\delta) + \kappa^T r.\end{aligned}$$

By choosing $T \le \frac{\log(r) + \log(1/\varepsilon(n,\delta))}{\log(1/\bar{\kappa})}$, we know $\bar{\kappa}^T r \le \varepsilon(n,\delta)$, hence

$$\|\theta_n^T - \theta^*\| \le \left(\frac{\bar{c}_5}{1-\bar{\kappa}} + 1\right) \varepsilon(n,\delta).$$

Take $\bar{C}_1$, $\bar{C}_2$ accordingly, we conclude the proof. □

B Proofs of Main Results

In this Appendix, we provide proofs for the results in the main text.

B.1 Proof of Lemma 1

We start from the following lemma:

Lemma 5. *Assume Assumption (W.1) holds, we have that*

$$f(\theta) - f(\theta^*) \ge \frac{c_1 \|\theta - \theta^*\|^{\alpha+2}}{(\alpha+1)(\alpha+2)}.$$

Proof. Consider $g(\theta) = f(\theta) - \frac{c_1\|\theta-\theta^*\|^{\alpha+2}}{(\alpha+1)(\alpha+2)}$. With Assumption (W.1), we know that

$$\nabla^2 g(\theta) = \nabla^2 f(\theta) - \frac{c_1}{(\alpha+1)(\alpha+2)} \left(\alpha(\alpha+2)\|\theta\|^{\alpha-2}\theta\theta^\top - (\alpha+2)\|\theta\|^\alpha I\right) \succeq 0,$$

as the operator norm of $\alpha(\alpha+2)\|\theta\|^{\alpha-2}\theta\theta^\top + (\alpha+2)\|\theta\|^\alpha I$ is less than $(\alpha+1)(\alpha+2)\|\theta\|^\alpha$. Meanwhile, we have that

$$\nabla g(\theta) = \nabla f(\theta) - \frac{c_1\|\theta\|^\alpha}{\alpha+1}(\theta - \theta^*).$$

As $\nabla f(\theta^*) = 0$, we know $\nabla g(\theta^*) = 0$, which means θ^* is the minimizer of g. Hence,

$$f(\theta^*) = g(\theta^*) \le g(\theta) = f(\theta) - \frac{c_1\|\theta - \theta^*\|^{\alpha+2}}{(\alpha+1)(\alpha+2)},$$

which means

$$f(\theta) - f(\theta^*) \ge \frac{c_1\|\theta - \theta^*\|^{\alpha+2}}{(\alpha+1)(\alpha+2)}.$$

As a consequence, we obtain the conclusion of Lemma 5. □

Now, we prove Lemma 1. Notice that

$$
\begin{aligned}
\|F(\theta)-\theta^*\|^2 =& \left\|\theta - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\nabla f(\theta) - \theta^*\right\|^2 \\
=&\|\theta-\theta^*\|^2 - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\langle \nabla f(\theta), \theta-\theta^*\rangle + \frac{\eta^2}{\lambda^2_{\max}(\nabla^2(f(\theta)))}\|\nabla f(\theta)\|^2 \\
=&\|\theta-\theta^*\|^2 - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\left(\langle \nabla f(\theta), \theta-\theta^*\rangle - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\|\nabla f(\theta)\|^2\right) \\
\leq&\|\theta-\theta^*\|^2 - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\left(f(\theta)-f(\theta^*) - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\|\nabla f(\theta)\|^2\right),
\end{aligned}
$$

where the last inequality is due to the convexity. With Assumption (W.1), we have that

$$
\begin{aligned}
\|\nabla f(\theta)\| =& \left\|\int_0^1 \nabla^2 f(\theta^* + t(\theta-\theta^*))(\theta-\theta^*)dt\right\| \\
\leq& \int_0^1 \left\|\nabla^2 f(\theta^* + t(\theta-\theta^*))(\theta-\theta^*)\right\| dt \\
\leq& \int_0^1 \lambda_{\max}(\nabla^2 f(\theta^* + t(\theta-\theta^*)))\|\theta-\theta^*\|dt \\
\leq& \int_0^1 c_2 t^\alpha \|(\theta-\theta^*)\|^\alpha \|\theta-\theta^*\|dt \\
\leq& \frac{c_2}{\alpha+1}\|\theta-\theta^*\|^{\alpha+1}.
\end{aligned}
$$

As $\eta \leq \frac{c_1^2}{2c_2^2} \leq \frac{c_1^2(\alpha+1)}{c_2^2(\alpha+2)}$, we have that

$$
\begin{aligned}
&\frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\left(f(\theta)-f(\theta^*) - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\|\nabla f(\theta)\|^2\right) \\
&\qquad\qquad\qquad \geq \frac{\eta}{c_2}\left(\frac{c_1}{(\alpha+1)(\alpha+2)} - \frac{\eta c_2^2}{c_1(\alpha+1)^2}\right)\|\theta-\theta^*\|^2.
\end{aligned}
$$

Hence, we find that

$$
\|F(\theta)-\theta^*\|^2 \leq \left(1 - \frac{\eta}{c_2}\left(\frac{c_1}{(\alpha+1)(\alpha+2)} - \frac{\eta c_2^2}{c_1(\alpha+1)^2}\right)\right)\|\theta-\theta^*\|^2.
$$

Take κ accordingly, we conclude the proof.

B.2 Proof of Lemma 2

Notice that

$$
\begin{aligned}
\|F_n(\theta)-F(\theta)\| =& \left\|\frac{\eta}{\lambda_{\max}(\nabla^2 f_n(\theta))}\nabla f_n(\theta) - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta))}\nabla f(\theta)\right\| \\
=&\eta\left\|\frac{\nabla f_n(\theta)\lambda_{\max}(\nabla^2 f(\theta)) - \nabla f(\theta)\lambda_{\max}(\nabla^2 f_n(\theta))}{\lambda_{\max}(\nabla^2 f_n(\theta))\lambda_{\max}(\nabla^2 f(\theta))}\right\| \\
\leq&\eta\left(\left\|\frac{\nabla f(\theta)(\lambda_{\max}(\nabla^2 f(\theta)) - \lambda_{\max}(\nabla^2 f_n(\theta)))}{\lambda_{\max}(\nabla^2 f_n(\theta))\lambda_{\max}(\nabla^2 f(\theta))}\right\| + \left\|\frac{\nabla f_n(\theta)-\nabla f(\theta)}{\lambda_{\max}(\nabla^2 f_n(\theta))}\right\|\right).
\end{aligned}
$$

For the term $\|\nabla f_n(\theta) - \nabla f(\theta)\|$, we have that

$$\begin{aligned}
\|\nabla f_n(\theta) - \nabla f(\theta)\| \leq& \|\nabla f_n(\theta^*) - \nabla f(\theta^*)\| \\
&+ \left\| \int_0^1 (\nabla^2 f_n(\theta^* + t(\theta - \theta^*)) - \nabla^2 f(\theta^* + t(\theta - \theta^*)))(\theta - \theta^*) dt \right\| \\
\leq& \int_0^1 \|(\nabla^2 f_n(\theta^* + t(\theta - \theta^*)) - \nabla^2 f(\theta^* + t(\theta - \theta^*)))(\theta - \theta^*)\| dt \\
\leq& \int_0^1 \|\nabla^2 f_n(\theta^* + t(\theta - \theta^*)) - \nabla^2 f(\theta^* + t(\theta - \theta^*))\|_{\mathrm{op}} \|\theta - \theta^*\| dt \\
\leq& \int_0^1 c_3 t^\gamma \epsilon(n, \delta) \|\theta - \theta^*\|^{\gamma+1} dt \\
=& \frac{c_3 \|\theta - \theta^*\|^{\gamma+1} \epsilon(n, \delta)}{\gamma + 1}.
\end{aligned}$$

Meanwhile, it's straightforward to show that

$$|\lambda_{\max}(\nabla^2 f_n(\theta)) - \lambda_{\max}(\nabla f(\theta))| \leq 3c_3 r^\gamma \epsilon(n, \delta).$$

Hence, we have that

$$\begin{aligned}
\|F_n(\theta) - F(\theta)\| \leq& \eta \left(\left\| \frac{\nabla f(\theta)(\lambda_{\max}(\nabla^2 f(\theta)) - \lambda_{\max}(\nabla^2 f_n(\theta)))}{\lambda_{\max}(\nabla^2 f_n(\theta)) \lambda_{\max}(\nabla^2 f(\theta))} \right\| + \left\| \frac{\nabla f_n(\theta) - \nabla f(\theta)}{\lambda_{\max}(\nabla^2 f_n(\theta))} \right\| \right) \\
\leq& \eta \left(\frac{3c_2 c_3 r^{\gamma+1-\alpha} \epsilon(n, \delta)}{(\alpha + 1)(c_1 r^\alpha - 3c_3 r^\gamma \epsilon(n, \delta)) c_1 r^\alpha} + \frac{c_3 r^\gamma \epsilon(n, \delta)}{(\gamma + 1)(c_1 r^\alpha - 3c_3 r^{\gamma+1} \epsilon(n, \delta))} \right).
\end{aligned}$$

As $r \geq \left(\frac{6c_3 \epsilon(n,\delta)}{c_1}\right)^{1/(\alpha-\gamma)}$, we can further have

$$\|F_n(\theta) - F(\theta)\| \leq \eta \left(\frac{6c_2 c_3}{(\alpha + 1)c_1^2} + \frac{2c_3}{(\gamma + 1)c_1} \right) r^{\gamma+1-\alpha} \epsilon(n, \delta).$$

Taking c_4 accordingly, we conclude the proof.

B.3 Proof of Theorem 1

Recall that for the radius of r_n in Lemma 2, we denote $r_n = \bar{C} \cdot \varepsilon(n, \delta)^{\frac{1}{\alpha - \gamma}}$. Without loss of generality, we assume $\|\theta_n^k - \theta^*\| > \left(\frac{c_4 \bar{C}^{\gamma+1-\alpha}}{1-\kappa} + 1\right) r_n$ holds for all $k < T$ where c_4 is the universal constant in Lemma 2, $T := C \log(1/\varepsilon(n, \delta)$ and C is some constant that will be chosen later; otherwise the conclusion of the theorem already holds.

We first show that, $\theta_n^k \in \mathbb{B}(\theta^*, r) \backslash \mathbb{B}(\theta^*, r_n)$ for all $k < T$. The inequality $\|\theta_n^k - \theta^*\| > r_n$ is direct from the hypothesis. Therefore, we only need to prove that $\|\theta_n^k - \theta^*\| \leq r$. Indeed, we have

$$\begin{aligned}
\|\theta_n^{k+1} - \theta^*\| =& \|F_n(\theta_n^k) - \theta^*\| \\
\leq& \|F_n(\theta_n^k) - F(\theta_n^k)\| + \|F(\theta_n^k) - \theta^*\| \\
\leq& \sup_{\theta \in \mathbb{B}(\theta^*, r) \backslash \mathbb{B}(\theta^*, r_n)} \|F_n(\theta) - F(\theta)\| + \|F(\theta_n^k) - \theta^*\| \\
\overset{(i)}{\leq}& c_4 r^{\gamma+1-\alpha} \varepsilon(n, \delta) + \kappa \|\theta_n^k - \theta^*\|
\end{aligned}$$

$$
\begin{aligned}
&\overset{(ii)}{\leq} c_4 \bar{C}^{\gamma+1-\alpha} \varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}} + \kappa r \\
&\overset{(iii)}{\leq} r
\end{aligned}
$$

with probability $1-\delta$ where the inequality (i) is due to Lemma 2 and c_4 is the universal constant in that lemma; the inequality (ii) is due to $r > r_n = \bar{C}\varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}}$ and $\gamma \leq \alpha$; the inequality (iii) is due to the assumption that n is sufficiently large such that $c_4 \bar{C}^{\gamma+1-\alpha} \varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}} \leq (1-\kappa)r$. As a consequence, we can guarantee that $\theta_n^k \in \mathbb{B}(\theta^*, r) \backslash \mathbb{B}(\theta^*, r_n)$ for all $k < T$.

Now, we would like to show that $\|\theta_n^T - \theta^*\| \leq \frac{2-\kappa}{1-\kappa} r_n$. Indeed, following the earlier argument, we find that

$$
\begin{aligned}
\|\theta_n^T - \theta^*\| \leq & \|F_n(\theta_n^{T-1}) - F(\theta_n^{T-1})\| + \|F(\theta_n^{T-1}) - \theta^*\| \\
\leq & \sup_{\theta \in \mathbb{B}(\theta^*, r) \backslash \mathbb{B}(\theta^*, r_n)} \|F_n(\theta) - F(\theta)\| + \kappa \|\theta_n^{T-1} - \theta^*\| \\
\leq & c_4 \cdot r_n^{\gamma+1-\alpha} \varepsilon(n,\delta) + \kappa \|\theta_n^{T-1} - \theta^*\| \\
= & c_4 \bar{C}^{\gamma+1-\alpha} \cdot \varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}} + \kappa \|\theta_n^{T-1} - \theta^*\|.
\end{aligned}
$$

By repeating the above argument T times, we finally obtain that

$$
\begin{aligned}
\|\theta_n^T - \theta^*\| \leq & c_4 \bar{C}^{\gamma+1-\alpha} \cdot \varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}} \left(\sum_{t=0}^{T-1} \kappa^t \right) + \kappa^T \|\theta_n^0 - \theta^*\| \\
\leq & \frac{c_4 \bar{C}^{\gamma+1-\alpha}}{1-\kappa} \varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}} + \kappa^T r.
\end{aligned}
$$

By choosing T such that $\kappa^T r \leq \varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}}$, which is equivalent to $T \geq \frac{\log(r) + \frac{1}{\alpha-\gamma}\log(1/\varepsilon(n,\delta))}{\log(1/\kappa)}$, we can guarantee that

$$
\|\theta_n^T - \theta^*\| \leq \left(\frac{c_4 \bar{C}^{\gamma+1-\alpha}}{1-\kappa} + 1 \right) \varepsilon(n,\delta)^{\frac{1}{\alpha-\gamma}}.
$$

As a consequence, we obtain the conclusion of the theorem.

B.4 Proof of Proposition 1

With convexity Cauchy-Schwartz inequality and Lemma 5, we have that

$$
\frac{c_1 \|\theta - \theta^*\|^{\alpha+2}}{(\alpha+1)(\alpha+2)} \leq f(\theta) - f(\theta^*) \leq \langle \nabla f(\theta), \theta - \theta^* \rangle \leq \|\nabla f(\theta)\| \|\theta - \theta^*\|.
$$

Hence, we know

$$
\|\nabla f(\theta)\| \geq \frac{c_1 \|\theta - \theta^*\|^{\alpha+1}}{(\alpha+1)(\alpha+2)}.
$$

Moreover, straightforward computation shows

$$
f(\theta - f(\theta)^*) \leq \frac{c_2 \|\theta - \theta^*\|^{\alpha+2}}{(\alpha+1)(\alpha+2)}.
$$

Hence, we know there exists a universal constant c depends on c_1, c_2, α, such tat

$$
\|\nabla f(\theta)\| \geq c(f(\theta) - f(\theta^*))^{1-\frac{1}{\alpha+2}}
$$

The following Lemma will be helpful for the proof.

Lemma 6 (Lemma 3.5 in [3])**.** *If f is β-smooth, then $\forall \theta_1, \theta_2 \in \mathbb{R}^d$, we have that*

$$f(\theta_1) - f(\theta_2) \leq \langle \nabla f(\theta_1), \theta_1 - \theta_2 \rangle - \frac{1}{2\beta} \|\nabla f(\theta_1) - \nabla f(\theta_2)\|^2.$$

Apply Lemma 6 for (θ_1, θ_2) and (θ_2, θ_1), we obtain the following corollary.

Corollary 3. *If f is β-smooth, then $\forall \theta_1, \theta_2 \in \mathbb{R}^d$, we have that*

$$\frac{1}{\beta} \|\nabla f(\theta_1) - \nabla f(\theta_2)\|^2 \leq \langle \nabla f(\theta_1) - \nabla f(\theta_2), x - y \rangle.$$

Notice that, if $\theta_1, \theta_2 \in \mathbb{B}(\theta^*, r)$, then $\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leq c_1 r^\alpha \|\theta_1 - \theta_2\|$, which means f is $c_1 r^\alpha$-smooth in $\mathbb{B}(\theta^*, r)$. With the standard step-size choice of smooth function optimization, we assume the step-size satisfies $0 < \eta < \frac{2}{c_1 r^\alpha}$, and define the effective step-size as $\frac{1}{\beta} := \eta(2 - c_1 r^\alpha \eta) > 0$ where $\beta > c_1 r^\alpha$. If $\theta^t_{\text{GD}} \in \mathbb{B}(\theta^*, r)$, we have that

$$\begin{aligned}
\|\theta^{t+1} - \theta^*\|^2 - \|\theta^t - \theta^*\|^2 =& \eta^2 \|\nabla f(\theta^t)\|^2 - 2\eta \langle \nabla f(\theta^t), \theta^t - \theta^* \rangle \\
\leq & -\frac{1}{\beta} \langle \nabla f(\theta^t), \theta^t - \theta^* \rangle \leq 0,
\end{aligned}$$

where the last inequality is due to Corollary 3. Hence, $\theta^{t+1}_{\text{GD}} \in \mathbb{B}(\theta^*, r)$. Furthermore, with Assumption (W.1), we have

$$\begin{aligned}
f(\theta^{t+1}) - f(\theta^t) \leq & \nabla f(\theta^t)^\top (\theta^{t+1} - \theta^t) + \frac{c_1 r^\alpha}{2} \|\theta^{t+1} - \theta^t\|^2 \\
=& -\frac{1}{2\beta} \|\nabla f(\theta^t)\|^2 \\
\leq & -\frac{c^2}{2\beta} (f(\theta^t) - f(\theta^*))^{2 - \frac{2}{\alpha+2}} \leq 0.
\end{aligned}$$

We then need the following lemma from [16]:

Lemma 7. *Given $\alpha > 0$, $\forall x \in [0, 1]$,*

$$\frac{1}{\alpha}(1 - x^\alpha) \geq x^\alpha (1 - x).$$

Proof. Consider the mapping $g : x \mapsto \frac{1}{\alpha}(x^\alpha - 1) - x^\alpha(1 - x)$.. We can see $g(0) = \frac{1}{\alpha}$ and $g(1) = 0$. Moreover,

$$\nabla g(x) = -(\alpha + 1)(x^{\alpha - 1} - x^\alpha) \leq 0,$$

which concludes the proof. □

Define $\delta(\theta^t) := f(\theta^t) - f(\theta^*)$, we have that

$$\begin{aligned}
\frac{1}{\delta(\theta^t)^{\frac{\alpha}{\alpha+2}}} =& \frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}} + \sum_{s=1}^{t-1} \left(\frac{1}{\delta(\theta^s)^{\frac{\alpha}{\alpha+2}}} - \frac{1}{\delta(\theta^{s+1})^{\frac{\alpha}{\alpha+2}}} \right) \\
=& \frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}} + \sum_{s=1}^{t-1} \frac{\frac{\alpha}{\alpha+2}}{\delta(\theta^{s+1})^{\frac{\alpha}{\alpha+2}}} \cdot \frac{\alpha+2}{\alpha} \cdot \left(1 - \left(\frac{\delta(\theta^{s+1})}{\delta(\theta^s)} \right)^{\frac{\alpha}{\alpha+2}} \right)
\end{aligned}$$

$$
\begin{aligned}
\geq & \frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}} + \sum_{s=1}^{t-1} \frac{\frac{\alpha}{\alpha+2}}{\delta(\theta^{s+1})^{\frac{\alpha}{\alpha+2}}} \cdot \left(\frac{\delta(\theta^{s+1})}{\delta(\theta^s)}\right)^{\frac{\alpha}{\alpha+2}} \left(1 - \frac{\delta(\theta^{s+1})}{\delta(\theta^s)}\right) \\
= & \frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}} + \sum_{s=1}^{t-1} \frac{\frac{\alpha}{\alpha+2}}{\delta(\theta^s)^{2-\frac{2}{\alpha+2}}} \cdot (\delta(\theta^s) - \delta(\theta^{s+1})) \\
\geq & \frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}} + \sum_{s=1}^{t-1} \frac{\frac{\alpha}{\alpha+2}}{\delta(\theta^s)^{2-\frac{2}{\alpha+2}}} \cdot \frac{c^2}{2\beta} (\delta(\theta^t))^{2-\frac{2}{\alpha+2}} \\
= & \frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}} + \sum_{s=1}^{t-1} \frac{c^2 \left(\frac{\alpha}{\alpha+2}\right)}{2\beta} \\
= & \frac{1}{\delta(\theta^1)^{\frac{\alpha}{\alpha+2}}} + \frac{c^2 \left(\frac{\alpha}{\alpha+2}\right)}{2\beta} \cdot (t-1).
\end{aligned}
$$

We can conclude that

$$
f(\theta^t) - f(\theta^*) \leq \left[\frac{1}{(f(\theta^1) - f(\theta^*))^{\frac{\alpha}{\alpha+2}}} + \frac{c^2 \left(\frac{\alpha}{\alpha+2}\right)}{2\beta} \cdot (t-1)\right]^{-\frac{\alpha+2}{\alpha}} \leq C(\eta \cdot t)^{-\frac{\alpha+2}{\alpha}},
$$

where C is some universal constant. With Lemma 5, we obtain that $\|\theta^t - \theta^*\| \leq c_0(\eta t)^{-1/\alpha}$ for some universal constant c_0. Then invoking Theorem 1 in [11] along with Lemma 2, we can obtain the desired conclusion.

C Proof of Generalized Linear Models

In this appendix, we provide the proof for the NormGD in generalized linear models.

C.1 Homogeneous assumptions

Based on the formulation of the population loss function $\mathcal{L}$ in equation (8), we have

$$
\begin{aligned}
&\nabla \mathcal{L}(\theta) = 2p(2p-1)!!(\theta - \theta^*)\|\theta - \theta^*\|^{2p-2}, \\
&\nabla^2 \mathcal{L}(\theta) = (2p(2p-1)!!)\|\theta - \theta^*\|^{2p-4} \left(\|\theta - \theta^*\|^2 I_d + (2p-4)(\theta - \theta^*)(\theta - \theta^*)^\top\right).
\end{aligned}
$$

Notice that, $\theta - \theta^*$ is an eigenvector of $\|\theta - \theta^*\|^2 I_d + (2p-4)(\theta - \theta^*)(\theta - \theta^*)^\top$ with eigenvalue $(2p-3)\|\theta - \theta^*\|^2$, and any vector that is orthogonal to $\theta - \theta^*$ (which forms a $d-1$ dimensional subspace) is an eigenvector of $\|\theta - \theta^*\|^2 I_d + (2p-4)(\theta - \theta^*)(\theta - \theta^*)^\top$ with eigenvalue $\|\theta - \theta^*\|^2$. Hence, we have that

$$
\begin{aligned}
\lambda_{\max}(\|\theta - \theta^*\|^2 I_d + (2p-4)(\theta - \theta^*)(\theta - \theta^*)^\top) =& (2p-3)\|\theta - \theta^*\|^2, \\
\lambda_{\min}(\|\theta - \theta^*\|^2 I_d + (2p-4)(\theta - \theta^*)(\theta - \theta^*)^\top) =& \|\theta - \theta^*\|^2,
\end{aligned}
$$

which shows that $\mathcal{L}(\theta)$ satisfies the homogeneous assumption.

C.2 Uniform concentration bound

The proof for the concentration bound (6) is in Appendix D.1 of [22]; therefore, it is omitted. We focus on proving the uniform concentration bounds (7) and (11) for the Hessian matrix $\nabla^2 \mathcal{L}_n(\theta)$ around the Hessian matrix $\nabla^2 \mathcal{L}(\theta)$ under both the strong and low signal-to-noise regimes. Indeed, we would like to show the following uniform concentration bound that captures both the bounds (7) and (11).

Lemma 8. *There exist universal constants C_1 and C_2 such that as long as $n \geq C_1 (d \log(d/\delta))^{2p}$ we obtain that*

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla^2 \mathcal{L}_n(\theta) - \nabla^2 \mathcal{L}(\theta)\|_{op} \leq C_2 \left((r + \|\theta^*\|)^{p-2} + (r + \|\theta^*\|)^{2p-2} \right) \sqrt{\frac{d + \log(1/\delta)}{n}}. \quad (19)$$

Proof of Lemma 8. Direct calculation shows that

$$\begin{aligned}
\nabla^2 \mathcal{L}_n(\theta) =& \frac{1}{n} \sum_{i=1}^n \left(p(X_i^\top \theta)^{2p-2} - p(p-1) Y_i (X_i^\top \theta)^{p-2} \right) X_i X_i^\top, \\
\nabla^2 \mathcal{L}(\theta) =& \mathbb{E} \left[p(X^\top \theta)^{2p-2} - p(p-1)(X^\top \theta^*)^p (X^\top \theta)^{p-2} X X^\top \right].
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\nabla^2 \mathcal{L}_n(\theta) - \nabla^2 \mathcal{L}(\theta) =& p \left(\frac{1}{n} \sum_{i=1}^n (X_i^\top \theta)^{2p-2} X_i X_i^\top - \mathbb{E}\left[(X^\top \theta)^{2p-2} X X^\top \right] \right) \\
& - p(p-1) \left(\frac{1}{n} \sum_{i=1}^n Y_i (X_i^\top \theta)^{p-2} X_i X_i^\top - \mathbb{E}\left[(X^\top \theta^*)^p (X^\top \theta)^{p-2} X X^\top \right] \right).
\end{aligned}$$

Using the triangle inequality with the operator norm, the above equation leads to

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla^2 \mathcal{L}_n(\theta) - \nabla^2 \mathcal{L}(\theta)\|_{\text{op}} \leq C(A_1 + A_2 + A_3), \quad (20)$$

where C is some universal constant and A_1, A_2, A_3 are defined as follows:

$$\begin{aligned}
A_1 =& \sup_{\theta \in \mathbb{B}(\theta^*, r)} \left\| \frac{1}{n} \sum_{i=1}^n (X_i^\top \theta)^{2p-2} X_i X_i^\top - \mathbb{E}\left[(X^\top \theta)^{2p-2} X X^\top \right] \right\|_{\text{op}}, \\
A_2 =& \sup_{\theta \in \mathbb{B}(\theta^*, r)} \left\| \frac{1}{n} \sum_{i=1}^n (Y_i - (X_i^\top \theta^*)^p)(X_i^\top \theta)^{p-2} X_i X_i^\top \right\|_{\text{op}}, \\
A_3 =& \sup_{\theta \in \mathbb{B}(\theta^*, r)} \left\| \frac{1}{n} \sum_{i=1}^n (X_i^\top \theta^*)^p (X_i^\top \theta)^{p-2} X_i X_i^\top - \mathbb{E}\left[(X^\top \theta^*)^p (X^\top \theta)^{p-2} X X^\top \right] \right\|_{\text{op}}. \quad (21)
\end{aligned}$$

With variational characterization of the operator norm and upper bound the norm of any $\theta \in \mathbb{B}(\theta^*, r)$ with $r + \|\theta^*\|$, we have

$$\begin{aligned}
A_1 &\leq (r + \|\theta^*\|)^{2p-2} T_1, \\
A_2 &\leq (r + \|\theta^*\|)^{p-2} T_2, \\
A_3 &\leq (r + \|\theta^*\|)^{p-2} T_3,
\end{aligned}$$

where the terms T_1, T_2, T_3 are defined as follows:

$$T_1 := \sup_{u\in\mathbb{S}^{d-1},\theta\in\mathbb{S}^{d-1}} \left| \frac{1}{n}\sum_{i=1}^n (X_i^\top\theta)^{2p-2}(X_i^\top u)^2 - \mathbb{E}\left[(X^\top\theta)^{2p-2}(X^\top u)^2\right] \right|$$
$$T_2 := \sup_{u\in\mathbb{S}^{d-1},\theta\in\mathbb{S}^{d-1}} \left| \frac{1}{n}\sum_{i=1}^n (Y_i - (X_i^\top\theta^*)^p)(X_i^\top\theta)^{p-2}(X_i^\top u)^2 \right|$$
$$T_3 := \sup_{u\in\mathbb{S}^{d-1},\theta\in\mathbb{S}^{d-1}} \left| \frac{1}{n}\sum_{i=1}^n (X_i^\top\theta^*)^p(X_i^\top\theta)^{p-2}(X_i^\top u)^2 - \mathbb{E}\left[(X^\top\theta^*)^p(X^\top\theta)^{p-2}(X^\top u)^2\right] \right|,$$

where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$.

Bound for T_2: With standard discretization arguments (e.g. Chapter 6 in [25]), let U be a 1/8-cover of $\mathbb{S}^{d-1}$ under $\|\cdot\|_2$ whose cardinality can be upper bounded by 17^d, we know

$$\sup_{u\in\mathbb{S}^{d-1}} \left| \frac{1}{n}\sum_{i=1}^n (Y_i - (X_i^\top\theta^*)^p)(X_i^\top\theta)^{p-2}(X_i^\top u)^2 \right| \le 2\sup_{u\in U} \left| \frac{1}{n}\sum_{i=1}^n (Y_i - (X_i^\top\theta^*)^p)(X_i^\top\theta)^{p-2}(X_i^\top u)^2 \right|$$

With a symmetrization argument, we know for any even integer $q \ge 2$,

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n (Y_i - (X_i^\top\theta^*)^p)(X_i^\top\theta)^{p-2}(X_i^\top u)^2\right)^q \le \mathbb{E}\left(\frac{2}{n}\sum_{i=1}^n \varepsilon_i(Y_i - (X_i^\top\theta^*)^p)(X_i^\top\theta)^{p-2}(X_i^\top u)^2\right)^q,$$

where $\{\varepsilon_i\}_{i\in[n]}$ is a set of i.i.d. Rademacher random variables. For a compact set Ω, define

$$\mathcal{R}(\Omega) := \sup_{\theta\in\Omega, p'\in[1,p]} \left| \frac{2}{n}\sum_{i=1}^n \varepsilon_i(Y_i - (X_i^\top\theta^*)^p)(X_i^\top\theta)^{p'-2}(X_i^\top u)^2 \right|,$$

and $\mathcal{N}(t)$ is a t-cover of $\mathbb{S}^{d-1}$ under $\|\cdot\|_2$. Then,

$$\begin{aligned}
\mathcal{R}(\mathbb{S}^{d-1}) &= \sup_{\theta\in\mathbb{S}^{d-1}, p'\in[2,p]} \left| \frac{2}{n}\sum_{i=1}^n \varepsilon_i(Y_i - (X_i^\top\theta^*)^p)(X_i^\top\theta)^{p'-2}(X_i^\top u)^2 \right| \\
&\le \sup_{\theta_t\in\mathcal{N}(t), \|\eta\|\le t, p'\in[2,p]} \left| \frac{2}{n}\sum_{i=1}^n \varepsilon_i(Y_i - (X_i^\top\theta^*)^p)(X_i^\top(\theta_t+\eta))^{p'-2}(X_i^\top u)^2 \right| \\
&\le \sup_{\theta_t\in\mathcal{N}(t), p'\in[2,p]} \left| \frac{4}{n}\sum_{i=1}^n \varepsilon_i(Y_i - (X_i^\top\theta^*)^p)(X_i^\top\theta_t)^{p'-2}(X_i^\top u)^2 \right| \\
&\quad + \max_{p'\in[2,p]} 3^{p'-2} \left| \frac{2}{n}\sum_{i=1}^n \varepsilon_i(X_i^\top\theta^*)^p(X_i^\top\eta)^{p'-2}(X_i^\top u)^2 \right| \\
&\le 2\mathcal{R}(\mathcal{N}(t)) + 3^{p-2}t\mathcal{R}(\mathbb{S}^{d-1}).
\end{aligned}$$

Take $t = 3^{-p+1}$, we have that $\mathcal{R}(\mathbb{S}^{d-1}) \le 3\mathcal{R}(\mathcal{N}(3^{-p+1}))$. We then move to the upper bound of $\mathcal{R}(\mathcal{N}(3^{-p+1}))$. With the union bound, for any $q \ge 1$ we have that

$$\sup_{\theta\in\mathbb{S}^{d-1}, p'\in[2,p]} \mathbb{E}\left[\left| \frac{2}{n}\sum_{i=1}^n \varepsilon_i(Y_i - (X_i^\top\theta^*)^p((X_i^\top\theta)^{p'-2}(X_i^\top u)^2 \right|^q\right]$$

$$
\begin{aligned}
&= \sup_{\theta \in \mathbb{S}^{d-1}, p' \in [2,p]} \int_0^\infty \mathbb{P}\left(\left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i (Y_i - (X_i^\top \theta^*)^p)(X_i^\top \theta)^{p'-2}(X_i^\top u)^2 \right|^q \geq \varepsilon \right) d\varepsilon \\
&\geq \sup_{\theta \in \mathcal{N}(3^{-p+1}), p' \in [2,p]} \int_0^\infty \mathbb{P}\left(\left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i (Y_i - (X_i^\top \theta^*)^p)(X_i^\top \theta)^{p'-2}(X_i^\top u)^2 \right|^q \geq \varepsilon \right) d\varepsilon \\
&\geq \frac{\sup_{p' \in [2,p]} \sum_{\theta \in \mathcal{N}(3^{-p+1})} \int_0^\infty \mathbb{P}\left(\left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i (Y_i - (X_i^\top \theta^*)^p)(X_i^\top \theta)^{p'-2}(X_i^\top u)^2 \right|^q \geq \varepsilon \right) d\varepsilon}{|\mathcal{N}(3^{-p+1})|} \\
&\geq \frac{\sup_{p' \in [2,p]} \int_0^\infty \mathbb{P}\left(\sup_{\theta \in \mathcal{N}(3^{-p+1})} \left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i (Y_i - (X_i^\top \theta^*)^p)(X_i^\top \theta)^{p'-2}(X_i^\top u)^2 \right|^q \geq \varepsilon \right) d\varepsilon}{|\mathcal{N}(3^{-p})|} \\
&\geq \frac{\int_0^\infty \mathbb{P}\left(\sup_{\theta \in \mathcal{N}(3^{-p+1}), p' \in [2,p]} \left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i (Y_i - (X_i^\top \theta^*)^p)(X_i^\top \theta)^{p'-2}(X_i^\top u)^2 \right|^q \geq \varepsilon \right) d\varepsilon}{p|\mathcal{N}(3^{-p+1})|} \\
&= \frac{\mathbb{E}[\mathcal{R}^q(\mathcal{N}(3^{-p+1}))]}{p|\mathcal{N}(3^{-p+1})|}.
\end{aligned}
$$

Hence, it's sufficient to consider $\mathbb{E}\left[\left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i (Y_i - (X_i^\top \theta^*)^p)(X_i^\top \theta)^{p'-2}(X_i^\top u)^2 \right|^q \right]$. We apply Khintchine's inequality [2], which guarantees that there is an universal constant C, such that for all $p' \in [2,p]$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i (Y_i - (X_i^\top \theta^*)^p)(X_i^\top \theta)^{p'-2}(X_i^\top u)^2 \right|^q \right] \\
&\qquad\qquad \leq \mathbb{E}\left[\left(\frac{Cq}{n^2} \sum_{i=1}^n (Y_i - (X_i^\top \theta^*)^p)^2 (X_i^\top \theta)^{2(p'-2)} (X_i^\top u)^4 \right)^{q/2} \right]
\end{aligned}
$$

To further upper bound the right hand side of the above equation, we consider the large deviation property of random variable $(Y_i - (X_i^\top \theta^*)^p)^2 (X_i^\top \theta)^{2(p'-2)} (X_i^\top u)^4$. It's straightforward to show that

$$
\begin{aligned}
\mathbb{E}\left[(Y_i - (X_i^\top \theta^*)^p)^2 (X_i^\top \theta)^{2(p'-2)} (X_i^\top u)^4 \right] &\leq (2p')^{p'}, \\
\mathbb{E}\left[\left((Y - (X_i^\top \theta^*)^p)^2 (X_i^\top \theta)^{2(p'-4)} (X_i^\top u)^4 \right)^{q/2} \right] &\leq (2p'q)^{p'q}.
\end{aligned}
$$

With Lemma 2 in [17], with probability at least $1-\delta$, we have

$$
\begin{aligned}
&\left| \frac{1}{n} \sum_{i=1}^n \left((Y - (X_i^\top \theta^*)^p)^2 (X_i^\top \theta)^{2(p'-2)} (X_i^\top u)^4 \right)^{q/2} - \mathbb{E}\left[(Y_i - (X_i^\top \theta^*)^p)^2 (X_i^\top \theta)^{2(p'-2)} (X_i^\top u)^4 \right] \right| \\
&\quad \leq (8p')^{p'} \sqrt{\frac{\log 4/\delta}{n}} + (2p' \log(n/\delta))^{p'} \frac{\log 4/\delta}{n}.
\end{aligned}
$$

Hence, we have that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^n (Y_i - (X_i^\top \theta^*)^p)^2 (X_i^\top \theta)^{2(p'-1)} (X_i^\top u)^4 \right)^{q/2} \right] \\
&\qquad \leq 2^{q/2} \left(\mathbb{E}\left[(Y_i - (X_i^\top \theta^*)^p)^2 (X_i^\top \theta)^{2(p'-2)} (X_i^\top u)^4 \right] \right)^{q/2}
\end{aligned}
$$

$$
\begin{aligned}
&+2^{q/2}\mathbb{E}\Bigg[\Bigg|\sum_{i=1}^{n}\Big((Y_i-(X_i^\top\theta^*)^p)^2(X_i^\top\theta)^{2(p'-2)}(X_i^\top u)^4\Big)^{q/2} \\
&-\mathbb{E}\left[(Y_i-(X_i^\top\theta^*)^p)^2(X_i^\top\theta)^{2(p'-2)}(X_i^\top u)^4\right]\Bigg|^{q/2}\Bigg] \\
\le&(4p')^{p'q}+2^{q/2}\int_0^\infty \mathbb{P}\Bigg[\Bigg|\sum_{i=1}^{n}\Big((Y_i-(X_i^\top\theta^*)^p)^2(X_i^\top\theta)^{2(p'-2)}(X_i^\top u)^4\Big)^{q/2} \\
&-\mathbb{E}\left[(Y_i-(X_i^\top\theta^*)^p)^2(X_i^\top\theta)^{2(p'-2)}(X_i^\top u)^4\right]\Bigg|\ge\lambda\Bigg]d\lambda^{q/2} \\
\le&(4p')^{p'q}+2^{q/2}q(p'+1) \\
&\cdot\int_0^1\delta\left((8p')^{(p')}\sqrt{\frac{\log 4/\delta}{n}}+\frac{(2p'\log(n/\delta))^{(p'+1)}}{n}\right)^{q/2}d\log(n/\delta) \\
\le&(4p')^{p'q}+C'p'q\left((32p')^{p'q/2}n^{-q/4})\Gamma(q/4)\right. \\
&\left.+(8p')^{(p'+1)q/2}n^{-q/2}\left((\log n)^{(p+1)q/2}+\Gamma((p'+1)q/2)\right)\right),
\end{aligned}
$$

where C' is a universal constant and $\Gamma(\cdot)$ is the Gamma function. Notice that

$$
\begin{aligned}
&\mathbb{E}\left[\left|\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i-(X_i^\top\theta^*)^p)(X_i^\top\theta)^{p-2}X_i^\top u-\mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-1}(X^\top u)^4]\right)\right|^q\right] \\
&\qquad\le\mathbb{E}[\mathcal{R}^q(\mathbb{S}^{d-1})] \\
&\qquad\le 3^q\mathbb{E}[\mathcal{R}(\mathcal{N}(3^{-p+1}))] \\
&\qquad\le 3^q p|\mathcal{N}(3^{-p+1})|\sup_{\theta\in\mathbb{S}^{d-1}p'\in[2,p]}\mathbb{E}\left[\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(Y_i-(X_i^\top\theta^*)^p)(X_i^\top\theta)^{p'-2}(X_i^\top u)^2\right|^q\right] \\
&\qquad\le p3^{dp+2d+q}\left(\frac{Cq}{n}\right)^{q/2}\Big((4p)^{pq}+2C'pq\,(32p)^{pq}\,n^{-q/4}\Gamma(q/4) \\
&\qquad+(8p)^{(p+1)q/2}n^{-q/2}\left((\log n)^{(p+1)q/2}+\Gamma((p+1)q/2)\right)\Big),
\end{aligned}
$$

for any $u\in U$. Eventually, with union bound, we obtain

$$
\begin{aligned}
&\left(\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}(Y_i-(X_i^\top\theta^*)^p)(X_i^\top\theta)^{p-2}X_iX_i^\top\right\|^q\right]\right)^{1/q} \\
&\le 2\left(\mathbb{E}\left[\sup_{u\in U}\left|\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i-(X_i^\top\theta^*)^p)(X_i^\top\theta)^{p-2}(X_i^\top u)^2-\mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-2}(X^\top u)^2]\right)\right|^q\right]\right)^{1/q} \\
&\le 2\left(\mathbb{E}\left[\sum_{u\in[U]}\left|\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-2}(X_i^\top u)^2-\mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-2}(X^\top u)^2]\right)\right|^q\right]\right)^{1/q} \\
&\le 2\cdot 17^{d/q}\sup_{u\in[U]}\mathbb{E}\left[\left|\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-2}(X_i^\top u)^2-\mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-2}(X^\top u)^2]\right)\right|^q\right]^{1/q} \\
&\le 2\cdot(17)^{d/q}\cdot 3^{\frac{dp+2d}{q}+1}\left[\sqrt{\frac{C_pq}{n}}+\left(\frac{C_pq}{n}\right)^{3/4}+\frac{C_p}{n}(\log n+q)^{(p+1)/2}\right],
\end{aligned}
$$

where C_p is a universal constant that only depends on p. Take $q = d(p+3) + \log(1/\delta)$ and use the Markov inequality, we get the following bound on the second term T_2 with probability $1-\delta$:

$$T_2 \leq c_1 \Bigg(\sqrt{\frac{d + \log(1/\delta)}{n}} + \frac{1}{n} \left(d + \log\left(\frac{n}{\delta}\right)\right)^{\frac{p+1}{2}} \Bigg). \tag{22}$$

Bound for T_1: With the same discretization argument, we have that

$$\begin{aligned}
&\sup_{u\in\mathbb{S}^{d-1}} \left| \frac{1}{n}\sum_{i=1}^n (X_i^\top \theta)^{2p-2}(X_i^\top u)^2 - \mathbb{E}\left[(X^\top\theta)^{2p-2}(X^\top u)^2\right] \right| \\
\leq& 2\sup_{u\in U} \left| \frac{1}{n}\sum_{i=1}^n (X_i^\top \theta)^{2p-2}(X_i^\top u)^2 - \mathbb{E}\left[(X^\top\theta)^{2p-2}(X^\top u)^2\right] \right|
\end{aligned}$$

We still apply a symmetrization argument and know that for any even integer q,

$$\begin{aligned}
&\mathbb{E}\left[\left| \left(\frac{1}{n}\sum_{i=1}^n (X_i^\top \theta)^{2p-2}(X_i^\top u)^2 - \mathbb{E}[(X^\top\theta)^{2p-2}(X^\top u)^2] \right) \right|^q \right] \\
&\qquad \leq \mathbb{E}\left[\left| \left(\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top \theta)^{2p-2}(X_i^\top u)^2 \right) \right|^q \right],
\end{aligned}$$

where $\{\varepsilon_i\}_{i\in[n]}$ is a set of i.i.d. Rademacher random variables. For a compact set Ω, define

$$\mathcal{R}(\Omega) := \sup_{\theta\in\Omega, p'\in[2,p]} \left| \frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top \theta)^{2p'-2}(X_i^\top u)^2 \right|,$$

and $\mathcal{N}(t)$ is a t-cover of $\mathbb{S}^{d-1}$ under $\|\cdot\|_2$. Then,

$$\begin{aligned}
\mathcal{R}(\mathbb{S}^{d-1}) =& \sup_{\theta\in\mathbb{S}^{d-1}, p'\in[2,p]} \left| \frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top \theta)^{2p'-2}(X_i^\top u)^2 \right| \\
\leq& \sup_{\theta_t\in\mathcal{N}(t), \|\eta\|\leq t, p'\in[2,p]} \left| \frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top (\theta_t+\eta))^{2p'-2}(X_i^\top u)^2 \right| \\
\leq& \sup_{\theta_t\in\mathcal{N}(t), p'\in[2,p]} \left| \frac{4}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top \theta_t)^{2p'-2}(X_i^\top u)^2 \right| \\
&+ \max_{p'\in[2,p]} 3^{2p'-2} \left| \frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top \eta)^{2p'-2}(X_i^\top u)^2 \right| \\
\leq& 2\mathcal{R}(\mathcal{N}(t)) + 3^{2p-2} t \mathcal{R}(\mathbb{S}^{d-1}).
\end{aligned}$$

Take $t = 3^{-2p+1}$, we have that $\mathcal{R}(\mathbb{S}^{d-1}) \leq 3\mathcal{R}(\mathcal{N}(3^{-2p+1}))$. We then move to the upper bound of $\mathcal{R}(\mathcal{N}(3^{-2p+1}))$. With the union bound, for any $q \geq 1$ we have that

$$\sup_{\theta\in\mathbb{S}^{d-1}, p'\in[2,p]} \mathbb{E}\left[\left| \frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top \theta)^{2p'-2}(X_i^\top u)^2 \right|^q \right]$$

$$
\begin{aligned}
&= \sup_{\theta\in\mathbb{S}^{d-1},p'\in[2,p]} \int_0^\infty \mathbb{P}\left(\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta)^{2p'-2}(X_i^\top u)^2\right|^q \geq \varepsilon\right) d\varepsilon \\
&\geq \sup_{\theta\in\mathcal{N}(3^{-2p+1}),p'\in[2,p]} \int_0^\infty \mathbb{P}\left(\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta)^{2p'-2}(X_i^\top u)^2\right|^q \geq \varepsilon\right) d\varepsilon \\
&\geq \frac{\sup_{p'\in[2,p]}\sum_{\theta\in\mathcal{N}(3^{-2p+1})}\int_0^\infty \mathbb{P}\left(\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta)^{2p'-2}(X_i^\top u)^2\right|^q \geq \varepsilon\right) d\varepsilon}{|\mathcal{N}(3^{-2p+1})|} \\
&\geq \frac{\sup_{p'\in[2,p]}\int_0^\infty \mathbb{P}\left(\sup_{\theta\in\mathcal{N}(3^{-2p+1})}\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta)^{2p'-2}(X_i^\top u)^2\right|^q \geq \varepsilon\right) d\varepsilon}{|\mathcal{N}(3^{-2p+1})|} \\
&\geq \frac{\int_0^\infty \mathbb{P}\left(\sup_{\theta\in\mathcal{N}(3^{-2p+1}),p'\in[2,p]}\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta)^{2p'-2}(X_i^\top u)^2\right|^q \geq \varepsilon\right) d\varepsilon}{p|\mathcal{N}(3^{-2p+1})|} \\
&= \frac{\mathbb{E}[\mathcal{R}^q(\mathcal{N}(3^{-2p+1}))]}{p|\mathcal{N}(3^{-2p+1})|}.
\end{aligned}
$$

Hence, it's sufficient to consider $\mathbb{E}\left[\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta)^{2p'-2}(X_i^\top u)^2\right|^q\right]$. We apply Khintchine's inequality [2], which guarantees that there is an universal constant C, such that for all $p' \in [2,p]$, we have

$$
\mathbb{E}\left[\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta)^{2p'-2}X_i^\top u\right|^q\right] \leq \mathbb{E}\left[\left(\frac{Cq}{n^2}\sum_{i=1}^n ((X_i^\top\theta)^{4p'-4)}(X_i^\top u)^4\right)^{q/2}\right]
$$

To further upper bound the right hand side of the above equation, we consider the large deviation property of random variable $(X_i^\top\theta)^{4p'-4}(X_i^\top u)^4$. It's straightforward to show that

$$
\begin{aligned}
\mathbb{E}\left[(X_i^\top\theta)^{4p'-4}(X_i^\top u)^4\right] &\leq (4p')^{p'}, \\
\mathbb{E}\left[\left((X_i^\top\theta)^{4p'-4)}(X_i^\top u)^4\right)^{q/2}\right] &\leq (4p'q)^{p'q}.
\end{aligned}
$$

With Lemma 2 in [17], with probability at least $1-\delta$, we have

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \left((X_i^\top\theta)^{4p'-4}(X_i^\top u)^4\right)^{q/2} - \mathbb{E}\left[(X_i^\top\theta)^{4p'-4}(X_i^\top u)^4\right]\right| \\
&\qquad \leq (16p')^{2p'}\sqrt{\frac{\log 4/\delta}{n}} + (4p'\log(n/\delta))^{2p'}\frac{\log 4/\delta}{n}.
\end{aligned}
$$

Hence, we have that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta)^{4p'-4}(X_i^\top u)^4\right)^{q/2}\right] \\
\leq& 2^{q/2}\left(\mathbb{E}\left[(X_i^\top\theta)^{4p'-4}(X_i^\top u)^4\right]\right)^{q/2} \\
&+ 2^{q/2}\mathbb{E}\left[\left|\sum_{i=1}^n \left((X_i^\top\theta)^{4p'-4}(X_i^\top u)^4\right)^{q/2} - \mathbb{E}\left[(X_i^\top\theta)^{4p'-4}(X_i^\top u)^4\right]\right|^{q/2}\right]
\end{aligned}
$$

$$
\begin{aligned}
\le&(8p')^{2p'q} \\
&+2^{q/2}\int_0^\infty \mathbb{P}\left[\left|\sum_{i=1}^n \left((X_i^\top\theta)^{4p'-2}(X_i^\top u)^4\right)^{q/2} - \mathbb{E}\left[(X_i^\top\theta)^{4p'-4}(X_i^\top u)^4\right]\right| \ge \lambda\right] d\lambda^{q/2} \\
\le&(8p')^{2p'q} + 2^{q/2}q(2p'+1) \\
&\cdot \int_0^1 \delta\left((16p')^{2p'}\sqrt{\frac{\log 4/\delta}{n}} + \frac{(4p'\log(n/\delta))^{(2p'+1)}}{n}\right)^{q/2} d\log(n/\delta) \\
\le&(8p')^{2p'q} + 2C'p'q\left((64p'))^{p')q}n^{-q/4}\right)\Gamma(q/4) \\
&+(16p'))^{(2p'+1)q/2}n^{-q/2}\left((\log n)^{(2p'+1)q/2} + \Gamma((2p'+1)q/2)\right)\Big),
\end{aligned}
$$

where C' is a universal constant and $\Gamma(\cdot)$ is the Gamma function. Notice that

$$
\begin{aligned}
\mathbb{E}&\left[\left|\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta)^{2p-2}(X_i^\top u)^2 - \mathbb{E}[(X^\top\theta)^{2p-2}(X^\top u)^2]\right)\right|^q\right] \\
&\le \mathbb{E}[\mathcal{R}^q(\mathbb{S}^{d-1})] \\
&\le 3^q\mathbb{E}[\mathcal{R}(\mathcal{N}(3^{-2p+1}))] \\
&\le 3^q p|\mathcal{N}(3^{-2p+1})| \sup_{\theta\in\mathbb{S}^{d-1}p'\in[2,p]} \mathbb{E}\left[\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta)^{2p'-2}(X_i^\top u)^2\right|^q\right] \\
&\le p3^{2pd+q}\left(\frac{Cq}{n}\right)^{q/2}\Big((8p)^{2pq} + 2C'pq\,(64p)^{pq}\,n^{-q/4}\Gamma(q/4) \\
&+(16p)^{(2p+1)q/2}n^{-q/2}\left((\log n)^{(2p+1)q/2} + \Gamma((2p+1)q/2)\right)\Big),
\end{aligned}
$$

for any $u \in U$. Eventually, with union bound, we obtain

$$
\begin{aligned}
&\left(\mathbb{E}\left[\left\|\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta)^{2p-2}X_iX_i^\top - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-1}XX^\top]\right)\right\|^q\right]\right)^{1/q} \\
&\quad\le 2\left(\mathbb{E}\left[\sup_{u\in U}\left|\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta)^{2p-2}(X_i^\top u)^2 - \mathbb{E}[(X^\top\theta)^{2p-2}(X^\top u)^2]\right)\right|^q\right]\right)^{1/q} \\
&\quad\le 2\left(\mathbb{E}\left[\sum_{u\in[U]}\left|\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta)^{2p-2}(X_i^\top u)^2 - \mathbb{E}[(X^\top\theta)^{2p-2}(X^\top u)^2]\right)\right|^q\right]\right)^{1/q} \\
&\quad\le 2\cdot 17^{d/q}\sup_{u\in[U]}\mathbb{E}\left[\left|\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta)^{2p-2}(X_i^\top u)^2 - \mathbb{E}[(X^\top\theta)^{2p-2}(X^\top u)^2]\right)\right|^q\right]^{1/q} \\
&\quad\le 6\cdot(17\cdot 3^{2p})^{d/q}\left[\sqrt{\frac{C_pq}{n}} + \left(\frac{C_pq}{n}\right)^{3/4} + \frac{C_p}{n}(\log n + q)^{(2p+1)/2}\right],
\end{aligned}
$$

where C_p is a universal constant that only depends on p. Take $q = d(2p+3) + \log(1/\delta)$ and use the Markov inequality, we get the following bound on the second term T_1 with probability $1-\delta$:

$$
T_1 \le c_2\Bigg(\sqrt{\frac{d+\log(1/\delta)}{n}} + \frac{1}{n}\left(d + \log\left(\frac{n}{\delta}\right)\right)^{\frac{2p+1}{2}}\Bigg). \tag{23}
$$

Bound for T_3: The same discretization argument yields

$$\begin{aligned}
&\sup_{u\in\mathbb{S}^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-2}(X_i^\top u)^2-\mathbb{E}\left[(X^\top\theta^*)^p(X^\top\theta)^{p-2}(X^\top u)^2\right]\right| \\
\le& 2\sup_{u\in U}\left|\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-2}(X_i^\top u)^2-\mathbb{E}\left[(X^\top\theta^*)^p(X^\top\theta)^{p-2}(X^\top u)^2\right]\right|.
\end{aligned}$$

And the same symmetrization argument shows for any even integer q,

$$\begin{aligned}
&\mathbb{E}\left[\left|\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-2}(X_i^\top u)^2-\mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-2}(X^\top u)^2]\right)\right|^q\right] \\
&\qquad\le \mathbb{E}\left[\left|\left(\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-2}(X_i^\top u)^2\right)\right|^q\right],
\end{aligned}$$

where $\{\varepsilon_i\}_{i\in[n]}$ is a set of i.i.d. Rademacher random variables. We then follow the proof strategy used in Section A.2 in [17]. For a compact set Ω, define

$$\mathcal{R}(\Omega):=\sup_{\theta\in\Omega,p'\in[2,p]}\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(X_i^\top\theta^*)^p(X_i^\top\theta)^{p'-2}(X_i^\top u)^2\right|,$$

and $\mathcal{N}(t)$ is a t-cover of $\mathbb{S}^{d-1}$ under $\|\cdot\|_2$. Then,

$$\begin{aligned}
\mathcal{R}(\mathbb{S}^{d-1})&=\sup_{\theta\in\mathbb{S}^{d-1},p'\in[1,p]}\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(X_i^\top\theta^*)^p(X_i^\top\theta)^{p'-2}(X_i^\top u)^2\right| \\
&\le \sup_{\theta_t\in\mathcal{N}(t),\|\eta\|\le t,p'\in[1,p]}\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(X_i^\top\theta^*)^p(X_i^\top(\theta_t+\eta))^{p'-2}(X_i^\top u)^2\right| \\
&\le \sup_{\theta_t\in\mathcal{N}(t),p'\in[1,p]}\left|\frac{4}{n}\sum_{i=1}^{n}\varepsilon_i(X_i^\top\theta^*)^p(X_i^\top\theta_t)^{p'-2}(X_i^\top u)^2\right| \\
&\quad+\max_{p'\in[1,p]}3^{p'-1}\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(X_i^\top\theta^*)^p(X_i^\top\eta)^{p'-2}(X_i^\top u)^2\right| \\
&\le 2\mathcal{R}(\mathcal{N}(t))+3^{p-2}t\mathcal{R}(\mathbb{S}^{d-1}).
\end{aligned}$$

Take $t=3^{-p+1}$, we have that $\mathcal{R}(\mathbb{S}^{d-1})\le 3\mathcal{R}(\mathcal{N}(3^{-p+1}))$. We then move to the upper bound of $\mathcal{R}(\mathcal{N}(3^{-p+1}))$. With the union bound, for any $q\ge 1$ we have that

$$\begin{aligned}
&\sup_{\theta\in\mathbb{S}^{d-1},p'\in[2,p]}\mathbb{E}\left[\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(X_i^\top\theta^*)^p(X_i^\top\theta)^{p'-2}(X_i^\top u)^2\right|^q\right] \\
&=\sup_{\theta\in\mathbb{S}^{d-1},p'\in[2,p]}\int_0^\infty\mathbb{P}\left(\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(X_i^\top\theta^*)^p(X_i^\top\theta)^{p'-2}(X_i^\top u)^2\right|^q\ge\varepsilon\right)d\varepsilon \\
&\ge\sup_{\theta\in\mathcal{N}(3^{-p+1}),p'\in[2,p]}\int_0^\infty\mathbb{P}\left(\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(X_i^\top\theta^*)^p(X_i^\top\theta)^{p'-2}(X_i^\top u)^2\right|^q\ge\varepsilon\right)d\varepsilon \\
&\ge\frac{\sup_{p'\in[1,p]}\sum_{\theta\in\mathcal{N}(3^{-p+1})}\int_0^\infty\mathbb{P}\left(\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(X_i^\top\theta^*)^p(X_i^\top\theta)^{p'-2}(X_i^\top u)^2\right|^q\ge\varepsilon\right)d\varepsilon}{|\mathcal{N}(3^{-p})|}
\end{aligned}$$

$$\geq \frac{\sup_{p'\in[1,p]} \int_0^\infty \mathbb{P}\left(\sup_{\theta\in\mathcal{N}(3^{-p+1})} \left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta^*)^p (X_i^\top\theta)^{p'-2}(X_i^\top u)^2\right|^q \geq \varepsilon\right) d\varepsilon}{|\mathcal{N}(3^{-p})|}$$

$$\geq \frac{\int_0^\infty \mathbb{P}\left(\sup_{\theta\in\mathcal{N}(3^{-p+1}),p'\in[1,p]} \left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta^*)^p (X_i^\top\theta)^{p'-2}(X_i^\top)^2 u\right|^q \geq \varepsilon\right) d\varepsilon}{p|\mathcal{N}(3^{-p+1})|}$$

$$= \frac{\mathbb{E}[\mathcal{R}^q(\mathcal{N}(3^{-p+1}))]}{p|\mathcal{N}(3^{-p+1})|}.$$

Hence, it's sufficient to consider $\mathbb{E}\left[\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta^*)^p (X_i^\top\theta)^{p'-2}(X_i^\top u)^2\right|^q\right]$. We apply Khintchine's inequality [2], which guarantees that there is an universal constant C, such that for all $p' \in [2, p]$, we have

$$\mathbb{E}\left[\left|\frac{2}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top\theta^*)^p (X_i^\top\theta)^{p'-2}(X_i^\top u)^2\right|^q\right] \leq \mathbb{E}\left[\left(\frac{Cq}{n^2}\sum_{i=1}^n (X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2p'-4}(X_i^\top u)^4\right)^{q/2}\right]$$

To further upper bound the right hand side of the above equation, we consider the large deviation property of random variable $(X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2p'-4}(X_i^\top u)^4$. It's straightforward to show that

$$\begin{aligned}
\mathbb{E}\left[(X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2p'-4}(X_i^\top u)^4\right] \leq& (2(p+p'))^{(p+p')},\\
\mathbb{E}\left[\left((X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2p'-4}(X_i^\top u)^4\right)^{q/2}\right] \leq& (2(p+p')q)^{(p+p')q}.
\end{aligned}$$

With Lemma 2 in [17], with probability at least $1-\delta$, we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \left((X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2p'-4}(X_i^\top u)^4\right)^{q/2} - \mathbb{E}\left[(X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2p'-4}(X_i^\top u)^4\right]\right|\\
&\quad\leq (8(p+p'))^{(p+p')}\sqrt{\frac{\log 4/\delta}{n}} + (2(p+p')\log(n/\delta))^{(p+p')}\frac{\log 4/\delta}{n}.
\end{aligned}$$

Hence, we have that

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n (X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2p'-4}(X_i^\top u)^4\right)^{q/2}\right]\\
\leq& 2^{q/2}\left(\mathbb{E}\left[(X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2p'-4}(X_i^\top u)^4\right]\right)^{q/2}\\
&+ 2^{q/2}\mathbb{E}\left[\left|\sum_{i=1}^n \left((X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2p'-4}(X_i^\top u)^4\right)^{q/2} - \mathbb{E}\left[(X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2p'-4}(X_i^\top u)^4\right]\right|^{q/2}\right]\\
\leq& (4(p+p'))^{(p+p')q}\\
&+ 2^{q/2}\int_0^\infty \mathbb{P}\left[\left|\sum_{i=1}^n \left((X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2p'-4}(X_i^\top u)^4\right)^{q/2} - \mathbb{E}\left[(X_i^\top\theta^*)^{2p}(X_i^\top\theta)^{2p'-4}(X_i^\top u)^4\right]\right| \geq \lambda\right] d\lambda^{q/2}\\
\leq& (4(p+p'))^{(p+p')q} + 2^{q/2}q(p+p'+1)\\
&\cdot \int_0^1 \delta\left((8(p+p'))^{(p+p')}\sqrt{\frac{\log 4/\delta}{n}} + \frac{(2(p+p')\log(n/\delta))^{(p+p'+1)}}{n}\right)^{q/2} d\log(n/\delta)
\end{aligned}$$

$$
\begin{aligned}
\leq & (4(p+p'))^{(p+p')q} + C'(p+p')q\Big((32(p+p'))^{(p+p')q/2} n^{-q/4})\Gamma(q/4) \\
& + (8(p+p'))^{(p+p'+1)q/2} n^{-q/2}\left((\log n)^{(p'+p+1)q/2} + \Gamma((p+p'+1)q/2)\right)\Big),
\end{aligned}
$$

where C' is a universal constant and $\Gamma(\cdot)$ is the Gamma function. Notice that

$$
\begin{aligned}
\mathbb{E} & \left[\left|\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-2}(X_i^\top u)^2 - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-2}(X^\top u)^2]\right)\right|^q\right] \\
& \leq \mathbb{E}[\mathcal{R}^q(\mathbb{S}^{d-1})] \\
& \leq 3^q\mathbb{E}[\mathcal{R}(\mathcal{N}(3^{-p+1}))] \\
& \leq 3^q p|\mathcal{N}(3^{-p+1})| \sup_{\theta\in\mathbb{S}^{d-1} p'\in[1,p]} \mathbb{E}\left[\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(X_i^\top\theta^*)^p(X_i^\top\theta)^{p'-2}(X_i^\top u)^2\right|^q\right] \\
& \leq 3^q p(3^{p+2})^d\left(\frac{Cq}{n}\right)^{q/2}\Big((16p)^{2pq} + 2C'pq\,(64p)^{pq}\,n^{-q/4}\Gamma(q/4) \\
& + (16p)^{(2p+1)q/2} n^{-q/2}\left((\log n)^{(2p+1)q/2} + \Gamma((2p+1)q/2)\right)\Big),
\end{aligned}
$$

for any $u \in U$. Eventually, with union bound, we obtain

$$
\begin{aligned}
& \left(\mathbb{E}\left[\left\|\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-2}X_iX_i^\top - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-2}XX^\top]\right)\right\|^q\right]\right)^{1/q} \\
& \leq 2\left(\mathbb{E}\left[\sup_{u\in U}\left|\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-2}(X_i^\top u)^2 - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-2}(X^\top u)^2]\right)\right|^q\right]\right)^{1/q} \\
& \leq 2\left(\mathbb{E}\left[\sum_{u\in[U]}\left|\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-2}(X_i^\top u)^2 - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-2}(X^\top u)^2]\right)\right|^q\right]\right)^{1/q} \\
& \leq 2\cdot 17^{d/q}\sup_{u\in[U]}\mathbb{E}\left[\left|\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\theta^*)^p(X_i^\top\theta)^{p-2}(X_i^\top u)^2 - \mathbb{E}[(X^\top\theta^*)^p(X^\top\theta)^{p-2}(X^\top u)^2]\right)\right|^q\right]^{1/q} \\
& \leq 6\cdot(17\cdot 3^{p+2})^{d/q}\left[\sqrt{\frac{C_pq}{n}} + \left(\frac{C_pq}{n}\right)^{3/4} + \frac{C_p}{n}(\log n + q)^{(2p+1)/2}\right],
\end{aligned}
$$

where C_p is a universal constant that only depends on p. Take $q = d(p+4) + \log(1/\delta)$ and use the Markov inequality, we get the following bound on the second term T_2 with probability $1-\delta$:

$$
T_3 \leq c_3\Bigg(\sqrt{\frac{d+\log(1/\delta)}{n}} + \frac{1}{n}\left(d + \log\left(\frac{n}{\delta}\right)\right)^{\frac{2p+1}{2}}\Bigg). \tag{24}
$$

Plugging the bounds (22), (23), and (24) to the bounds (20) and (21), and use the condition that $n \geq C_1(d\log(d/\delta))^{2p}$ we obtain the conclusion of the lemma. $\square$

D Proof of Gaussian Mixture Models

In this appendix, we provide the proof for the NormGD in Gaussian mixture models.

D.1 Homogeneous assumptions

Direct calculation shows that

$$\nabla^2 \bar{\mathcal{L}}(\theta) = \frac{1}{\sigma^2}\left(I_d - \frac{1}{\sigma^2}\mathbb{E}\left(XX^\top \mathrm{sech}^2\left(\frac{X^\top\theta}{\sigma^2}\right)\right)\right).$$

To simplify the calculation, we perform a change of coordinates via an orthogonal matrix R, such that $R\theta = \|\theta\| e_1$ where e_1 is the first canonical basis in dimension d. By denoting $V = \frac{RX}{\sigma}$, we have $V = (V_1, \cdots, V_d) \sim \mathcal{N}(0, I_d)$, and we can rewrite the above equation as

$$\nabla^2 \bar{\mathcal{L}}(\theta) = \frac{1}{\sigma^2}\left(I_d - \mathbb{E}_V\left(VV^\top \mathrm{sech}^2\left(\frac{V_1\|\theta\|}{\sigma}\right)\right)\right).$$

The matrix $A = \mathbb{E}_V\left(VV^\top \mathrm{sech}^2\left(\frac{V_1\|\theta\|}{\sigma}\right)\right)$ is a diagonal matrix with $A_{11} = \mathbb{E}_{V_1}\left[V_1^2 \mathrm{sech}^2\left(\frac{V_1\|\theta\|}{\sigma}\right)\right]$, $A_{ii} = \mathbb{E}_{V_1}\left[\mathrm{sech}^2\left(\frac{V_1\|\theta\|}{\sigma}\right)\right]$ for all $2 \le i \le d$. An application of $\mathrm{sech}^2(x) \ge 1 - x^2$ shows:

$$\begin{aligned}
A_{11} &= \mathbb{E}_{V_1}\left[V_1^2 \mathrm{sech}^2\left(\frac{V_1\|\theta\|}{\sigma}\right)\right] \ge \mathbb{E}_{V_1}\left[V_1^2\left(1 - \frac{V_1^2\|\theta\|^2}{\sigma^2}\right)\right] = 1 - \frac{3\|\theta\|^2}{\sigma^2}, \\
A_{ii} &= \mathbb{E}_{V_1}\left[\mathrm{sech}^2\left(\frac{V_1\|\theta\|}{\sigma}\right)\right] \ge \mathbb{E}_{V_1}\left[\left(1 - \frac{V_1^2\|\theta\|^2}{\sigma^2}\right)\right] = 1 - \frac{\|\theta\|^2}{\sigma^2}, \quad \forall\, 2 \le i \le d.
\end{aligned}$$

This implies that

$$\lambda_{\max}(\nabla^2 \bar{\mathcal{L}}(\theta)) \le \frac{3\|\theta\|^2}{\sigma^2}.$$

Another application of $\mathrm{sech}^2(x) \le 1 - x^2 + \frac{2}{3}x^4$ shows

$$\begin{aligned}
A_{11} &= \mathbb{E}_{V_1}\left[V_1^2 \mathrm{sech}^2\left(\frac{V_1\|\theta\|}{\sigma}\right)\right] \le \mathbb{E}_{V_1}\left[V_1^2\left(1 - \frac{V_1^2\|\theta\|^2}{\sigma^2} + \frac{2V_1^4\|\theta\|^4}{3\sigma^4}\right)\right] = 1 - \frac{3\|\theta\|^2}{\sigma^2} + \frac{10\|\theta\|^4}{\sigma^4}, \\
A_{ii} &= \mathbb{E}_{V_1}\left[\mathrm{sech}^2\left(\frac{V_1\|\theta\|}{\sigma}\right)\right] \le \mathbb{E}_{V_1}\left[\left(1 - \frac{V_1^2\|\theta\|^2}{\sigma^2} + \frac{2V_1^4\|\theta\|^4}{3\sigma^4}\right)\right] = 1 - \frac{\|\theta\|^2}{\sigma^2} + \frac{2\|\theta\|^4}{\sigma^4},
\end{aligned}$$

for all $2 \le i \le d$. When $\|\theta\| \le \frac{\sigma}{2}$, we have that $\frac{\|\theta\|^2}{\sigma^4} \le \frac{1}{4}$, and hence

$$\begin{aligned}
A_1 1 &\le 1 - \frac{3\|\theta\|^2}{\sigma^2} + \frac{10\|\theta\|^4}{\sigma^4} \le 1 - \frac{\|\theta\|^2}{2\sigma^2} \\
A_i i &\le 1 - \frac{\|\theta\|^2}{\sigma^2} + \frac{2\|\theta\|^4}{\sigma^4} \le\le 1 - \frac{\|\theta\|^2}{2\sigma^2}, \quad \forall\, 2 \le i \le d.
\end{aligned}$$

Hence, we have that

$$\lambda_{\min}(\nabla^2 \bar{L}(\theta)) \ge \frac{\|\theta\|^2}{2\sigma^2},$$

which concludes the proof.

D.2 Uniform Concentration Bounds for Mixture Models

See Corollary 4 in [1] for the proof of the uniform concentration result between $\nabla \bar{\mathcal{L}}_n(\theta)$ and $\nabla \bar{\mathcal{L}}(\theta)$ in equation (15) for the strong signal-to-noise regime. Now, we prove the uniform concentration bounds between $\nabla^2 \bar{\mathcal{L}}_n(\theta)$ and $\nabla \bar{\mathcal{L}}(\theta)$ in equations (15) and (18) for both the strong signal-to-noise and low signal-to-noise regimes. It is sufficient to prove the following lemma.

Lemma 9. *There exist universal constants C_1 and C_2 such that as long as $n \geq C_1 d \log(1/\delta)$ we obtain that*

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla^2 \mathcal{L}_n(\theta) - \nabla^2 \mathcal{L}(\theta)\|_{op} \leq C_2(\|\theta^*\| + \sigma^2)\sqrt{\frac{d + \log(1/\delta)}{n}}. \tag{25}$$

Proof of Lemma 9. For the sample log-likelihood function of the Gaussian mixture model, direct calculation shows that

$$\bar{\mathcal{L}}_n(\theta) = \frac{\|\theta\|^2}{2\sigma^2} - \frac{1}{n}\sum_{i=1}^n \log\left(\exp\left(-\frac{X_i^\top \theta}{\sigma^2}\right) + \exp\left(\frac{X_i^\top \theta}{\sigma^2}\right)\right) - \log(2(\sqrt{2\pi})^d \sigma^d).$$

Therefore, we find that

$$\nabla \bar{\mathcal{L}}_n(\theta) = \frac{\theta}{\sigma^2} - \frac{1}{n\sigma^2}\sum_{i=1}^n X_i \tanh\left(\frac{X_i^\top \theta}{\sigma^2}\right),$$
$$\nabla^2 \bar{\mathcal{L}}_n(\theta) = \frac{1}{\sigma^2}\left(I_d - \frac{1}{n\sigma^2}\sum_{i=1}^n X_i X_i^\top \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right)\right),$$

where $\text{sech}^2(x) = \frac{4}{(\exp(-x)+\exp(x))^2}$ for all $x \in \mathbb{R}$.

For the population log-likelihood function, we have

$$\nabla^2 \bar{\mathcal{L}}(\theta) = \frac{1}{\sigma^2}\left(I_d - \frac{1}{\sigma^2}\mathbb{E}\left(XX^\top \text{sech}^2\left(\frac{X^\top \theta}{\sigma^2}\right)\right)\right).$$

Therefore, we obtain that

$$\nabla^2 \bar{\mathcal{L}}_n(\theta) - \nabla^2 \bar{\mathcal{L}}(\theta) = \frac{1}{\sigma^4}\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right) - \mathbb{E}\left(XX^\top \text{sech}^2\left(\frac{X^\top \theta}{\sigma^2}\right)\right)\right).$$

Use the variational characterization of operator norm, it's sufficient to consider

$$T = \sup_{u \in \mathbb{S}^{d-1}, \theta \in \mathbb{B}(\theta^*, r)} \left|\frac{1}{n}\sum_{i=1}^n (X_i^\top u)^2 \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right) - \mathbb{E}\left((X_i^\top u)^2 \text{sech}^2\left(\frac{X^\top \theta}{\sigma^2}\right)\right)\right|.$$

With a standard discretization argument (e.g. Chapter 6 in [25]), let U be a 1/8-cover of $\mathbb{S}^{d-1}$ under $\|\cdot\|_2$ whose cardinality can be upper bounded by 17^d, we have that

$$\begin{aligned}
&\sup_{u \in \mathbb{S}^{d-1}} \left|\frac{1}{n}\sum_{i=1}^n (X_i^\top u)^2 \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right) - \mathbb{E}\left((X_i^\top u)^2 \text{sech}^2\left(\frac{X^\top \theta}{\sigma^2}\right)\right)\right| \\
\leq &2 \sup_{u \in U} \left|\frac{1}{n}\sum_{i=1}^n (X_i^\top u)^2 \text{sech}^2\left(\frac{X_i^\top \theta}{\sigma^2}\right) - \mathbb{E}\left((X_i^\top u)^2 \text{sech}^2\left(\frac{X^\top \theta}{\sigma^2}\right)\right)\right|.
\end{aligned}$$

Hence, we can focus on the error on the fixed u. With a symmetrization argument, for $\lambda > 0$, we have that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda \sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{1}{n}\sum_{i=1}^n (X_i^\top u)^2 \mathrm{sech}^2\left(\frac{X_i^\top\theta}{\sigma^2}\right) - \mathbb{E}\left((X_i^\top u)^2\mathrm{sech}^2\left(\frac{X^\top\theta}{\sigma^2}\right)\right)\right|\right)\right] \\
\le&\mathbb{E}\left[\exp\left(\sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top u)^2\mathrm{sech}^2\left(\frac{X_i^\top\theta}{\sigma^2}\right)\right|\right)\right],
\end{aligned}$$

where $\{\varepsilon_i\}$ is a set of i.i.d Rademacher random variable. Use the inequality $0 \le \mathrm{sech}(x) \le 1, \forall x$, we have

$$\mathbb{E}\left[\exp\left(\sup_{\theta\in\mathbb{B}(\theta^*,r)}\left|\frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top u)^2\mathrm{sech}^2\left(\frac{X_i^\top\theta}{\sigma^2}\right)\right|\right)\right] \le \mathbb{E}\left[\exp\left(\left|\frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top u)^2\right|\right)\right].$$

Notice that, $X_i^\top u \sim \frac{1}{2}\mathcal{N}(u^\top\theta^*, \sigma^2) + \frac{1}{2}\mathcal{N}(-u^\top\theta^*, \sigma^2)$ and $\mathbb{E}[(X^\top u)^2] = (u^\top\theta^*)^2 + \sigma^2$. When $t < \frac{1}{2\sigma^2}$, we also have that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(t(X_i^\top u)^2\right)\right] \\
&\quad=\frac{1}{2}\mathbb{E}_{x\sim\mathcal{N}(u^\top\theta^*,\sigma^2)}[\exp(tx^2)] + \frac{1}{2}\mathbb{E}_{x\sim\mathcal{N}(-u^\top\theta^*,\sigma^2)}[\exp(tx^2)] \\
&\quad=\mathbb{E}_{x\sim\mathcal{N}(u^\top\theta^*,\sigma^2)}[\exp(tx^2)] \\
&\quad=(1-2t\sigma^2)^{-1/2}\exp\left(\frac{t}{1-2\sigma^2 t}(u^\top\theta^*)^2\right).
\end{aligned}$$

With some calculus, we know when $|t| \le \frac{1}{4\sigma^2}$,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(t((X_i^\top u)^2 - ((u^\top\theta^*)^2+\sigma^2))\right)\right] \\
&\quad=(1-2t\sigma^2)^{-1/2}\exp\left(\frac{2\sigma^2t^2}{1-2\sigma^2t}(u^\top\theta^*)^2 - t\sigma^2\right) \\
&\quad\le\exp(4t^2\sigma^2(u^\top\theta^*)^2 + 4t^2\sigma^4).
\end{aligned}$$

Furthermore, as ε_i is independent of $X_i^\top u$, we have that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(t\varepsilon_i(X_i^\top u)^2\right)\right] \\
=&\frac{1}{2}\left[\mathbb{E}\left[\exp\left(t\varepsilon_i(X_i^\top u)^2\right)\right] + \mathbb{E}\left[\exp\left(-t\varepsilon_i(X_i^\top u)^2\right)\right]\right] \\
\le&\exp(4t^2\sigma^2(u^\top\theta^*)^2 + 4t^2\sigma^4)\left(\frac{1}{2}\exp(t((u^\top\theta^*)^2+\sigma^2)) + \frac{1}{2}\exp(-t((u^\top\theta^*)^2+\sigma^2))\right) \\
\le&\exp\left(4t^2\sigma^2(u^\top\theta^*)^2 + t^2\left((u^\top\theta^*)^2+\sigma^2\right)^2\right) \\
\le&\exp\left(3t^2\left((u^\top\theta^*)^2+\sigma^2\right)^2\right) \\
\le&\exp\left(3t^2\left(\|\theta^*\|^2+\sigma^2\right)^2\right),
\end{aligned}$$

Substitute back, we have that $\forall |\lambda| \le \frac{n}{8\sigma^2}$

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(\left|\frac{2\lambda}{n}\sum_{i=1}^{n}\varepsilon_i (X_i^\top u)^2\right|\right)\right] \\
&\qquad \le \mathbb{E}\left[\exp\left(\frac{2\lambda}{n}\sum_{i=1}^{n}\varepsilon_i (X_i^\top u)^2\right)\right] + \mathbb{E}\left[\exp\left(-\frac{2\lambda}{n}\sum_{i=1}^{n}\varepsilon_i (X_i^\top u)^2\right)\right] \\
&\qquad \le 2\exp\left(\frac{12\lambda^2}{n}(\|\theta^*\|^2+\sigma^2)^2\right),
\end{aligned}
$$

where in the last inequality we use the fact that $e^x + e^{-x} \le 2e^{x^2}$. With standard Chernoff method and a union bound of $u \in U$, we know there exist universal constants C_1 and C_2, such that the following inequality

$$
\sup_{u\in U}\left|\frac{1}{n}\sum_{i=1}^{n}(X_i^\top u)^2 \mathrm{sech}^2\left(\frac{X_i^\top\theta}{\sigma^2}\right) - \mathbb{E}\left((X_i^\top u)^2\mathrm{sech}^2\left(\frac{X^\top\theta}{\sigma^2}\right)\right)\right| \le C_2(\|\theta^*\|^2+\sigma^2)\sqrt{\frac{d+\log 1/\delta}{n}}
$$

holds with probability at least $1-\delta$ as long as $n \ge C_1(d+\log 1/\delta)$. This concludes the proof. $\square$